**AFRL-IF-RS-TR-2002-318**
**Final Technical Report**
**January 2003**

# INTELLIGENCE ANALYST ASSOCIATE (IAA)-CYC KNOWLEDGE EXTRACTION

**Veridian Engineering**

*APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.*

**AIR FORCE RESEARCH LABORATORY**
**INFORMATION DIRECTORATE**
**ROME RESEARCH SITE**
**ROME, NEW YORK**

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2002-318 has been reviewed and is approved for publication.

APPROVED: Paul M. Engelhart

PAUL M. ENGELHART
Project Engineer

FOR THE DIRECTOR:

JOSEPH CAMERA, Chief
Information & Intelligence Exploitation Division
Information Directorate

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 074-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

**1. AGENCY USE ONLY (Leave blank)**

**2. REPORT DATE**
JANUARY 2003

**3. REPORT TYPE AND DATES COVERED**
Final Jan 01 – May 02

**4. TITLE AND SUBTITLE**
INTELLIGENCE ANALYST ASSOCIATE (IAA)-CYC KNOWLEDGE EXTRACTION

**5. FUNDING NUMBERS**
C - F30602-99-D-0050/TASK 0010
PE - 62702F
PR - 459E
TA - IA
WU - 01

**6. AUTHOR(S)**
Jeannette Neal, Benjamin Rode, Chris Crowner, and Dave Gunning

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Veridian Engineering
4455 Genesee Street
PO Box 400
Buffalo New York 14225

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Air Force Research Laboratory/IFEA
32 Brooks Road
Rome New York 13441-4114

**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
AFRL-IF-RS-TR-2002-318

**11. SUPPLEMENTARY NOTES**

AFRL Project Engineer: Paul M. Engelhart/IFEA/(315) 330-4477/ Paul.Engelhart@rl.af.mil

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** ***(Maximum 200 Words)***

The objective of this project was to assess the feasibility of leveraging the capabilities and strengths of the Intelligence Analyst Associate (IAA) and the Cyc Knowledge Base (KB) in order to help alleviate the textual data overload that intelligence analysts experience. IAA has capabilities for processing large volumes of unstructured text, extracting information relevant to intelligence analysts, such as entities (people, organizations, locations, dates, and times) and simple events (subject, verbs, and objects), storing the extracted information in a structured database, and enabling the use of visualization and analysis tools. The Cyc KB is a formalized representation of a vast quantity of fundamental human knowledge (facts, rules, of thumb, and heuristics) and consists of terms and assertions which relate those terms. By leveraging the Cyc KB, significant capabilities were exploited that greatly benefited the IAA and its end users. These included the ability to represent domain dependent facts in the Cyc KB to identify, classify, and specify knowledge concerning relevant entities as well as the ability to represent rules in the KB and use the Cyc KB inference engine to allow information to be derived from identified entities and entity classifications. A "plug-in, plug-out" system framework was developed that served as the processing framework and testbed for the information extraction components, similar to the framework used for the IAA system. One of the main goals was aimed at reducing the time it takes to run a document through the IAA-Cyc system. The prototype system developed under a previous effort processed documents at a rate of 1.5 minutes per sentence. The new system processes documents at a rate of four seconds per document in which a document is typically comprised of 30-50 sentences.

**14. SUBJECT TERMS**
Information Extraction, Cyc Knowledge Base, Ontological Engineering, Intelligence Analysts, Coreference Resolution, Lexicon Development

**15. NUMBER OF PAGES**
76

**16. PRICE CODE**

| | | | |
|---|---|---|---|
| **17. SECURITY CLASSIFICATION OF REPORT** | **18. SECURITY CLASSIFICATION OF THIS PAGE** | **19. SECURITY CLASSIFICATION OF ABSTRACT** | **20. LIMITATION OF ABSTRACT** |
| UNCLASSIFIED | UNCLASSIFIED | UNCLASSIFIED | UL |

# TABLE OF CONTENTS

# TABLE OF FIGURES

# TABLE OF TABLES

# IAA-CYC Knowledge Extraction

## Final Technical Report

## 1 Introduction

The objective of this project was to assess the feasibility of leveraging the combined capabilities of the Intelligence Analyst Associate (IAA), an information extraction (IE) system, and the Cyc Knowledge Base (KB), a very large knowledge base, for monitoring domains of interest to intelligence analysts.

The goal of IAA is to help alleviate the textual data overload that intelligence analysts experience. IAA has capabilities for processing large volumes of unstructured text, extracting information relevant to intelligence analysts, such as entities and simple events, storing the extracted information in a structured database, and enabling the use of analysis & visualization (A&V) tools.

However, IAA needs the ability to perform further, more intelligent processing, using the context of the documents and that of the analysts' persistent knowledge bases or "bodies of knowledge" (BOKs) to automatically generate new information/knowledge and add this new knowledge to the analysts' BOKs in their domains of interest.

This IAA-Cyc Knowledge Extraction project is a second phase follow-on to the earlier project entitled "Leveraging Cyc for IAA". We will refer to the earlier project as "IAA-Cyc 1" and to this current project as "IAA-Cyc 2" for short. See the IAA-Cyc 1 Final Technical Report for more information on the earlier project.

This report comprises the Final Technical Report for this project entitled "IAA-Cyc Knowledge Extraction". Section 2 lists the referenced documents. Section 3 presents the driving problems and project goals, Section 4 provides a brief overview of IAA and the Cyc KB, Section 5 presents an overview of our technical approach, Section 6 summarizes the project accomplishments, Section 7 provides more detailed information on technical approach and accomplishments, Section 8 summarizes lessons learned and future directions, and Section 9 provides a list of acronyms.

# 2 Referenced Documents

The following is a list of relevant documents that were referenced within this Report or influenced the design of the system.

1. Allen, Kenneth W., Krumel, Glenn, Pollack, Jonathan D., *China's Air Force Enters the 21st Century*, RAND, 1995.

2. Cycorp, Inc., Cycorp web site providing information on the Cyc Knowledge Base and other knowledge based products at: http://www.cyc.com

3. Defense Advanced Research Projects Agency, *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, 25-27 August 1993, Baltimore, MD.

4. Defense Advanced Research Projects Agency, *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 6-8 November 1995, Columbia, MD.

5. Defense Advanced Research Projects Agency, *Proceedings of the Seventh Message Understanding Conference (MUC-7),* 29 April – 1 May 1998, Fairfax, VA.

6. Defense Advanced Research Projects Agency, *Proceedings of the TIPSTER Phase III 24-Month Workshop*, 3-15 October 1998, Baltimore, MD.

7. Lappin, S. and Leass, H., "A Syntactically Based Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics 20*, 1994, pp. 535-561.

8. Levin, B., *English Verb Classes and Alternations*, The University of Chicago Press, 1993.

9. Mulvenon, James C., *Professionalization of the Senior Chinese Officer Corps*, RAND, 1997.

10. National Institute of Standards and Technology (NIST), Automatic Content Extraction (ACE) Program Website, http://www.itl.nist.gov/iaui/894.01/tests/ace/index.htm

11. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., *A Comprehensive Grammar of the English Language*, Longman, 1985.

12. Veridian Engineering, *Intelligence Analyst Associate Software User Manual*, September 2000.

13. Veridian Engineering, Veridian Knowledge Management Internet site with information on analyst support tools/systems developed by Veridian such as IAA and the Document Content Analysis and Retrieval System (DCARS): http://www.dcars.com

14. Veridian Engineering, *Leveraging Cyc for IAA Final Technical Report,* AFRL-IF-RS-TR-2002-2, January 2002.

# 3 Driving Problems and Project Goals

## 3.1 Problems Driving the Project Objectives

The problems that drove the program objectives are based on discussions with analysts at the National Air Intelligence Center (NAIC) and the Joint Warfare Analysis Center (JWAC). The driving problems include:

- Analysts are plagued by information overload, especially the large volume of text documents and message traffic that they must examine in order to find and extract relevant information.
- Analysts cannot afford to miss information that impacts their analyses.
- Analysts need tools that focus on specialized information.
- Analysts require precise and reliable extracted information.
- Analysts have difficulty in converting and organizing extracted information into a form or tool that will support their analysis activities.
- Analysts do not have enough control over the information stored and manipulated by some of their tools/systems such as IAA.

## 3.2 Project Goals

The high-level goals of this IAA-Cyc 2 Project were to:

- Automatically populate analysts' bodies of knowledge (BOKs) or information level database tables from information extracted from text documents, especially unstructured prose text. Table 1 below illustrates an example type of table that the IAA-Cyc software would be designed to fill. This table holds information on a People's Liberation Army (PLA) person.
- Focus on extracting information on persons, organizations, equipment, and facilities.
- Take a domain-independent and domain-portable approach to information extraction to the extent possible.
- Use technical approaches that support extensibility of the software capabilities.
- Use the most appropriate technological approach for the implementation of each of the different component software capabilities.

In the area of software functionality, the high-level goals were to develop high priority software capabilities such as to perform the following:

- Identify persons, organizations, geopolitical entities, dates/times that are missed by trained statistical-based "off the shelf" (OTS) software components for named entity identification.
- Perform coreference resolution.
- Normalize extracted expressions for persons and organizations.
- Extract attributes and relations for identified entities.

- Infer or derive new information based upon what necessarily or likely follows from the information already extracted. Such new information might include attributes and relations for identified entities.
- Identify the actors, actions, and affected for a small class of events.
- Determine the conformity or consistency of extracted information when compared to that which was expected or already known.

**Table 1 Example analyst's BOK table that IAA-Cyc software would be designed to fill**

| **PERSON: Hengmei Huang** | | | | | | |
|---|---|---|---|---|---|---|
| **#** | **Title** | **Job Position** | **Main Org.** | **Admin. Unit** | **Time Period** | **Address** |
| 1 | Captain to Col. | Deputy Commander | PLAAF | Air Group or Squadron | Including part of 1976 | ? , China |
| 2 | Captain to BGen | ? | PLAAF | Suborganization of 7th Air Army | Sometime in 1984 | Guangxi, China |
| 3 | BGen to MGen | Commander | PLAAF | Command Post | Including February 1991 | Shanghai, China |
| 4 | MGen | Commander | PLAAF | Command Post | Including March 1992 | Shanghai, China |
| 5 | MGen | Deputy Commander | PLA | 8th National People's Congress | Starting early 1993, ending mid 1993 | ? , China |
| 6 | MGen | Commander | PLAAF | Chengdu MR Air Force | Mid 1993 thru at least part of 1994 | Chengdu, China |
| 7 | MGen | Deputy Commander | PLA | Chengdu MR | Including February 1995 | Chengdu, China |

The follow table presents the prioritized requirements or tasks for the project.

**Table 2 Prioritized Project Requirements/Tasks**

| PRIORITY | REQUIREMENT/TASK DESCRIPTION |
|---|---|
| High | Extract the names of specialized organizations and suborganizations |
| High | Extract the following attributes of persons:<br>• Name<br>• Aliases<br>• Titles<br>• Ranks<br>• Positions<br>• Professions<br>• Affiliation |
| High | Extract the following attributes of organizations:<br>• Name<br>• Aliases<br>• Type - Military, including subtypes<br>• Address |
| High | Extract a set of relationships that are an intersection of:<br>• Relationships expressed in sample NAIC documents<br>• Relationships involving the target domain of the Chinese military<br>• Relationships of special interest to JWAC analysts |
| High | Extract and determine the following meta-information:<br>• Time period during which the extracted information is true<br>• Document identifier<br>• Text offsets within document<br>• Judgment as to consistency and expectedness of information<br>• Confidence measure associated with the extracted information |
| High | Resolve coreferences between person and organization references |
| High | Determine normal forms for extracted persons and organizations |
| High | Load extracted information and meta-information into the IBOK |
| High | Judge extracted information as inconsistent or consistent and expected or unexpected |
| High | Derive information that necessarily follows from extracted information |
| High | Display extracted information in context of the document |
| High | Display information loaded into the IBOK sorted by person, organization/group, and conformity judgment |
| High | Display the basis for conformity judgments |
| High | Provide users ability to update the information in the database |
| Medium | Provide users with control over conformity checking |

| | |
|---|---|
| Medium | Derive information that likely follows from extracted information |
| Medium | Provide users with control over information to be derived |
| Low | Extract the names of specialized groups and facilities |
| Low | Extract the following attributes of persons:<br>• Gender<br>• Education<br>• Nationality<br>• Age<br>• Birth date<br>• Death date<br>• Address |
| Low | Extract the following attributes of organizations:<br>• Type - Political and other subtypes<br>• Nationality<br>• Age<br>• Established date<br>• Ending date |
| Low | Extract and determine the following meta-information:<br>• Who reported the information<br>• Time/date the information was reported<br>• Basis for judgment of consistency and expectedness<br>• Source of the information<br>• Source of the confidence measure |
| Low | Provide a utility for adding word list files to the lexicon |
| Low | Provide a utility for generalizing examples that express attributions |
| Low | Provide a utility for testing expression patterns |

Figure 1 below illustrates the types of data structures used to hold information about persons, organizations, countries, and job positions. Analogous structures are used for other entity types. The data structures essentially consist of slot-value pairs. The slots may represent attributes such as name, type, and gender for which the filler would be a data type such as a text string or number. Additionally, some slots may represent attributes, such as affiliation, spouse, or residence, whose values are links or pointers to other data structures representing other persons, organizations, etc. These links represent relationships between the different entities. The figure illustrates a link representing a relationship between a person and his/her job position, and indirectly to the organization within which the job position exists.

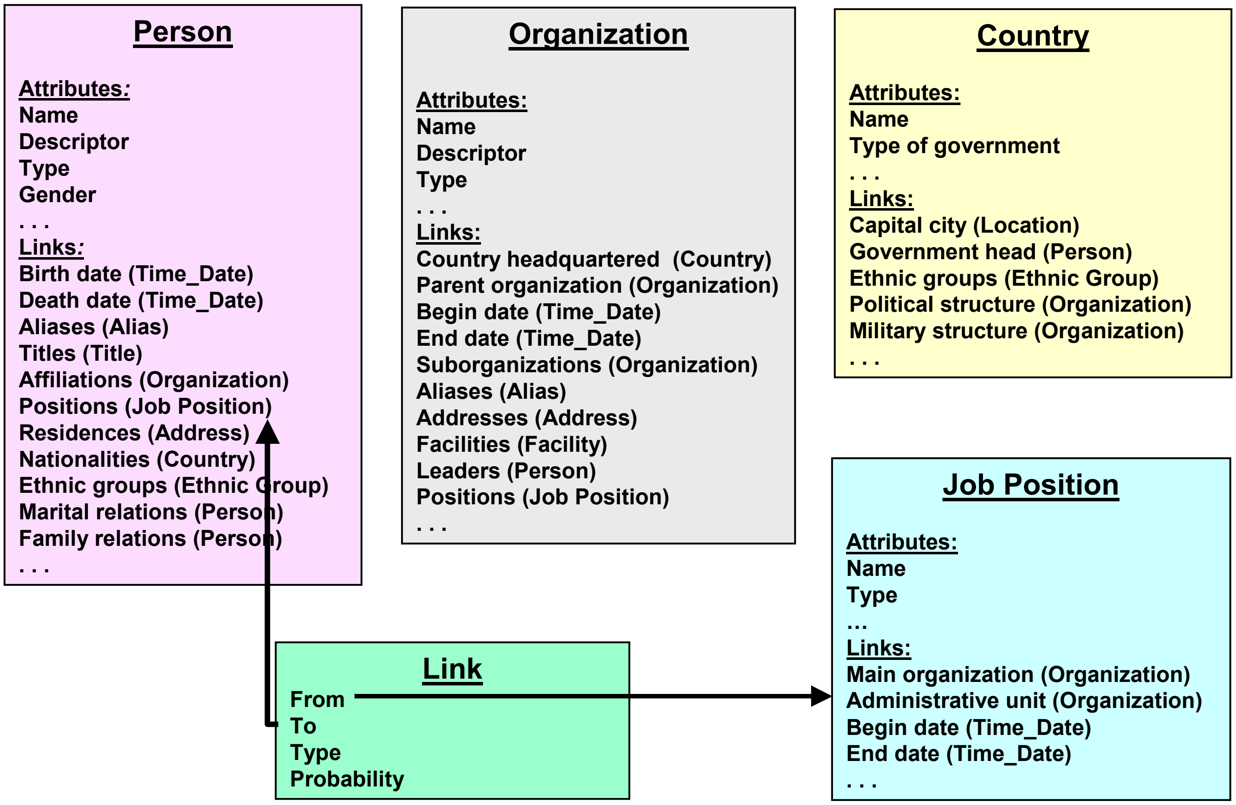


**Figure 1 IAA-Cyc uses techniques that support the representation (modeling) of entities and their attributes as well as links (relationships) between entities**

## 3.3 Project Scope

As in the previous IAA-Cyc 1 project, the following sources were used to help determine the candidate attributes and relationships to be targeted for automated extraction and insertion into an analyst's BOK as part of this project:

- Sample data, namely documents and messages, provided by analysts at NAIC and the JWAC.
- The NAIC Dynamic Information Operations Decision Environment (DIODE) Model and Database.
- Cyc Knowledge Base (KB).

Based on this study and consultation with the Government, the targeted types of attributes and relationships were selected and are listed below. These attributes and relationships were targeted for automatic extraction. Work on extracting the first five attribute types listed below was begun as part of IAA-Cyc Phase 1. This Phase 2 effort extended the accomplishments of Phase 1 and also developed new capabilities for attribute and relationship extraction.

- Names (including aliases).
- Positions (present and past).
- Military ranks (present and past).
- Branch of military service.
- Billet/military addresses.
- Relationships between persons, between organizations, and between persons and organizations including supervisor-supervisee, person with whom another person worked, and the organization for which a person worked.
- Time periods during which the attributes or relationships held.

It was agreed to continue to include the Chinese military as a domain of interest as was done in the IAA-Cyc 1 project, but to also exploit knowledge about generic classes of attributes, such as positions and titles, to be able to apply the software to automatically extract information across domains.

The following sources of information on the Chinese military were used in IAA-Cyc 1 and IAA-Cyc 2:

1. Allen, K.W., Krumel, G., Pollack, J.D., *China's Air Force Enters the 21st Century*, RAND, 1995.

2. Mulvenon, J.C., *Professionalization of the Senior Chinese Officer Corps*, RAND, 1997.

## 3.4 Project Milestones Achieved

Project milestones that were achieved include three Technical Interchange Meetings (TIMs) and software demonstrations, including a demonstration for the Scientific Advisory Board (SAB) during their visit to AFRL Rome. The following list presents the milestones:

- The Kickoff TIM was held on 8 February 2001 at AFRL Rome Research Site. The main purpose of the Kickoff Meeting was to:
  - Discuss program objectives, solution approaches and program evaluation.
  - Discuss and agree on a prioritization of requirements/tasks.
  - Discuss and finalize the program plan and schedule.

- The mid-term TIM was held on 13 September 2001 at AFRL Rome Research Site. The main purpose of the TIM Meeting was to:
  - Review accomplishments to date for the IAA-Cyc 2 Program.
  - Review and refine the plans, schedule, and directions for the IAA-Cyc 2 Program for the remaining contract performance period.
  - Provide a demonstration of IAA-Cyc 2 software prototype.

- The IAA-Cyc 2 software prototype was demonstrated at the Air Force Scientific Advisory Board (SAB) visit to AFRL Rome Research Site on 6 November 2001. The demonstration was presented at the poster session held at AFRL Rome for the SAB visit.

- The final TIM was held on 21 March 2002 at AFRL Rome Research Site. The purpose of the meeting was to:
  - Review accomplishments for the IAA-Cyc 2 Program.
  - Provide a demonstration of IAA-Cyc 2 software prototype.
  - Discuss future directions for the IAA-Cyc Program.

# 4 IAA and Cyc KB System Overviews

## 4.1 IAA Overview

IAA performs extraction of entities and simple events from high volume document repositories or feeds. IAA accepts ASCII documents from any source of text-based information: message traffic, reports, or open source text. IAA first applies a Text Zoner to locate the relevant parts of documents and messages and filter out the extraneous material from a document/message such as page breaks, headers, and footers. IAA then extracts the names of entities such as people, organizations, locations, dates and times. IAA also extracts shallow events in the form of subject, verb, direct and indirect objects. The extracted information is automatically loaded into a structured database for search and analysis.

In the A&V area, IAA provides a suite of eight (8) tools for analysts to use. These tools are summarized below:

- The Query Tool enables the user to create, edit, and execute queries that search the IAA database of information extracted from the documents/messages.
- The Statistics Tool enables the user to view information about the occurrence of single terms or phrases in a data set retrieved from the IAA database. The occurrence data is provided for each of the fields of the set of retrieved records.
- The Data Browser provides tabular visual displays of data sets retrieved from the IAA database. The Browser, for example, enables the user to view a dynamic table displaying the participants in simple events along with the location and date/time of the events, if available.
- The Document Browser enables the user to view and read the full text of any document in the IAA database, and view the location of the extracted information in the context of the full document/message.
- The Timeline Tool provides temporally-based visualizations of data sets retrieved from the IAA database. Each item in the data set (e.g., event) is represented by an icon on the timeline display with an associated descriptive text phrase and an associated horizontal bar that illustrates the duration or extent of the event or activity represented by the icon.
- The Geographic Display Tool provides geographical visualizations of data sets retrieved from the IAA database. The Geographic Display Tool displays icons for the data items on a map overlay display, placed appropriately to illustrate the location attribute of each item.
- The Topic Areas Tool enables analysts to save IAA database queries in a flexible and extensible hierarchical tree of folders that represent domains and topics of interest. Saved queries and the folder hierarchy are represented graphically using icons. Queries may be moved, copied, renamed, edited, run and displayed from within the tool. In this way, the tool provides a centralized topical organization of an analyst's work in IAA.

- The IAA Concept Domain Tool enables the analyst to define conceptual domain areas for which he/she is responsible, define different forms of the questions that he/she is tasked to answer, and edit the concept domain information to develop it over time. The purpose of the Concept Domain Tool is to enable the analyst-user to more quickly find and discover information on topics of interest and to enable the analyst to better control the precision of his/her search.

For more information on IAA, contact the AFRL Rome Research Site Program Manager.

## 4.2 Cyc KB Overview

The Cyc knowledge base (KB) is a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. The medium of representation is the formal language CycL. The KB consists of terms, which constitute the vocabulary of CycL, and assertions which relate those terms. These assertions include both simple ground assertions and rules. Cyc is not a frame-based system. Instead, the Cyc team thinks of the KB as a sea of assertions, with each assertion being no more “about” one of the terms involved than another.

The Cyc KB is divided into many (currently hundreds of) “microtheories”, each of which is essentially a bundle of assertions that share a common set of assumptions. Some microtheories are focused on a particular domain of knowledge, a particular level of detail, a particular interval in time, etc. The microtheory mechanism allows Cyc to independently maintain assertions which are *prima facie* contradictory, and enhances the performance of the Cyc system by focusing the inferencing process.

At the present time, the Cyc KB contains tens of thousands of terms and several dozen hand-entered assertions about or involving each term. New assertions are continually added to the KB by human knowledge enterers. The aforementioned numbers do not include (i) non-atomic terms such as predicates that express relationships between entities, nor (ii) the vast number of assertions added to the KB by Cyc itself as a product of the inferencing process.

The Cyc inference engine performs general logical deduction (including modus ponens, modus tolens, and universal and existential quantification), with AI's well-known named inference mechanisms (inheritance, automatic classification, etc.) as special cases. Cyc performs best-first search over a proof-space using a set of proprietary heuristics, and uses microtheories to optimize inferencing by restricting search domains.

For more information on the Cyc Knowledge Base and other knowledge based products, visit the Cycorp web site at: http://www.cyc.com

## 5 Technical Approach Overview

The high level design concept for the IAA-Cyc system is illustrated in the figure below. The main processing steps include the following:

- Text Zoning is the identification of the various parts of a message or document (e.g., header, addressee list, source, title, body) as well as extraneous items such as page breaks, headers, and footers.
- Information Identification is the recognition of text segments comprising expressions for items such as entities, entity attributes, relationships, and simple events. Semantic categorization is applied to assign semantic types or categories to the text expressions.
- Coreference Resolution is the determination as to which expressions refer to the same entities.
- Normalization is the conversion of text expressions into standard expressions for the entities or concepts; normalization was applied to identified text segments expressing the entity names (e.g., "Senator Clinton," "Clinton," and "Hillary" would all be mapped into a standard name such as "Hillary R. Clinton").
- Attribution & Relations Identification assigns extracted attributes with the entity with which they should be associated and identifies relationships among the entities.
- Information Inference refers to the process of inferring items of information from the extracted information that were not expressed explicitly in the text.
- The resulting extracted and inferred information is loaded the analyst's BOK database.

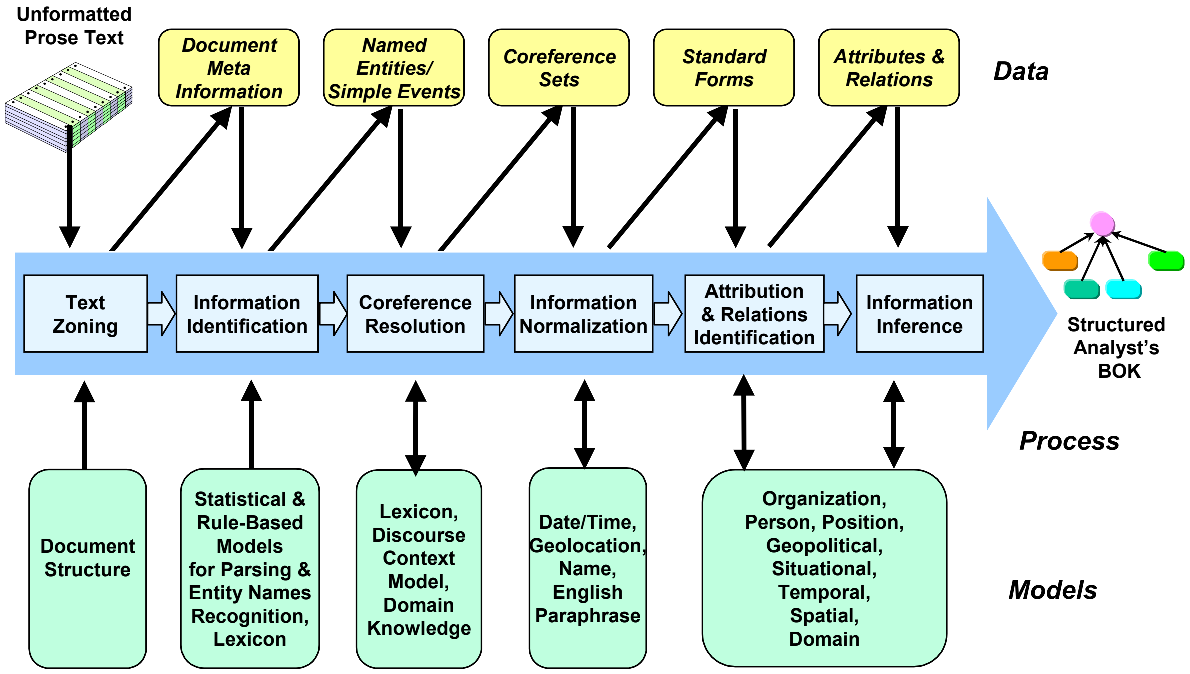

Unformatted Prose Text
Document Meta Information
Named Entities/ Simple Events
Coreference Sets
Standard Forms
Attributes & Relations
Data
Text Zoning
Information Identification
Coreference Resolution
Information Normalization
Attribution & Relations Identification
Information Inference
Structured Analyst's BOK
Process
Document Structure
Statistical & Rule-Based Models for Parsing & Entity Names Recognition, Lexicon
Lexicon, Discourse Context Model, Domain Knowledge
Date/Time, Geolocation, Name, English Paraphrase
Organization, Person, Position, Geopolitical, Situational, Temporal, Spatial, Domain
Models

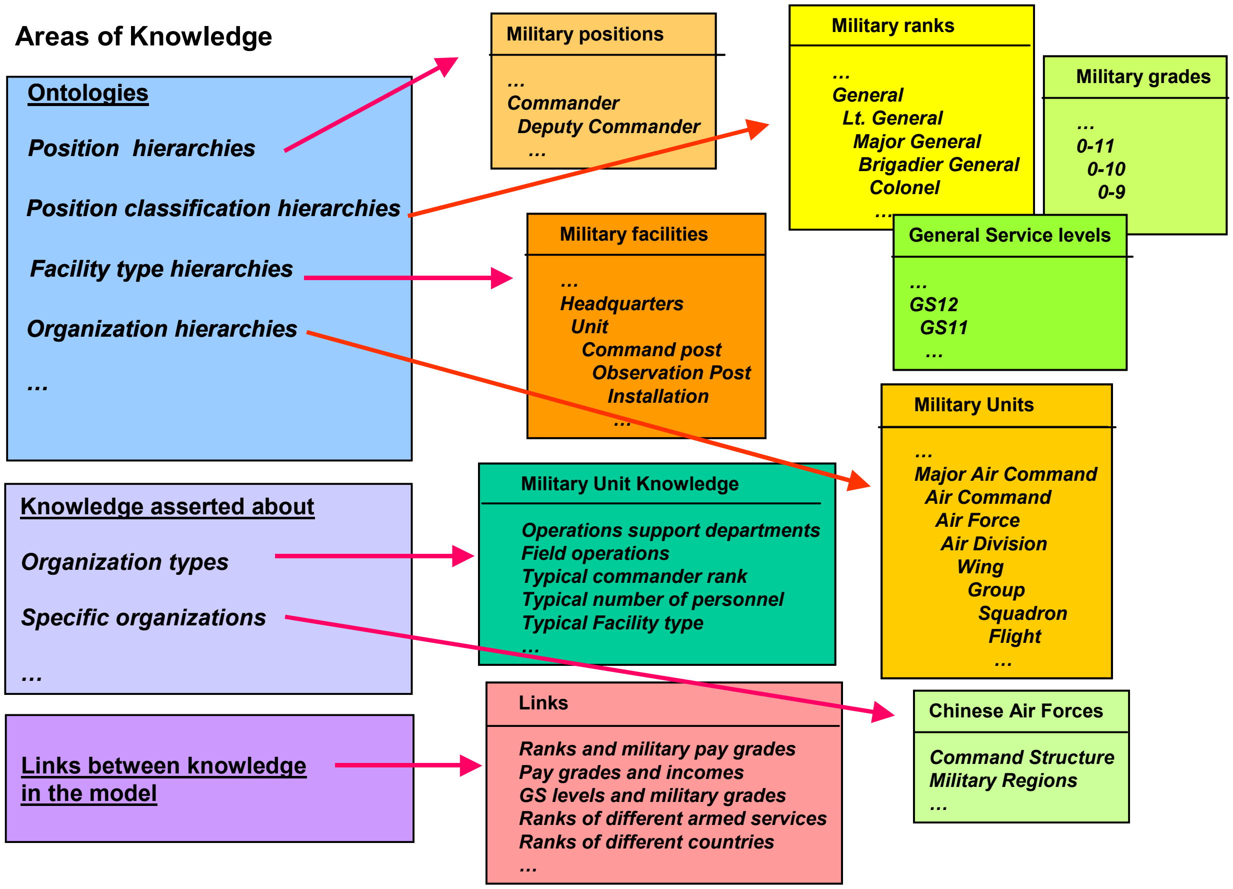
**Figure 2 The high level design concept for the IAA-Cyc system**

## 5.1 Leveraging the Cyc KB

The Cyc KB provides significant capabilities that can be leveraged to the benefit of IAA and its end users. The capabilities that were exploited in this project include:

- The ability to represent domain dependent facts in the Cyc KB to identify, classify and specify knowledge concerning relevant entities.
- The ability to represent rules in the KB and use the Cyc KB inference engine to allow information to be derived from identified entities and entity classifications.
- The ability to represent attributes of entities and their classifications.
- The ability to represent entity types and relations between the types.
- The ability to use microtheories for the representation of contexts.
- The ability to make use of ontological knowledge representation, permitting:
    - Different levels of generality in analysis allowing various degrees of domain independence to be maintained.
    - The ability to exploit inheritance and thereby gain benefits such as economy in the statement of rules.
    - Use of the existing wealth of knowledge previously developed and implemented in the Cyc KB, including both general common-sense knowledge and more domain-specific specialized knowledge in relevant areas.

The figure below illustrates some of the knowledge areas represented and used in the IAA-Cyc system. The figure indicates some of the ontologies used and the types of entities represented. These ontologies include military positions, ranks, and facilities. Links (relations) between the different entity types were also represented. Example relations include the relationship between a person and his/her position, as well as the relation between a person and the organization with which the person is affiliated.



Areas of Knowledge

**Ontologies**

*Position hierarchies*

*Position classification hierarchies*

*Facility type hierarchies*

*Organization hierarchies*

*…*

**Knowledge asserted about**

*Organization types*

*Specific organizations*

*…*

**Links between knowledge in the model**

Military positions

*…*
*Commander*
*Deputy Commander*
*…*

Military ranks

*…*
*General*
*Lt. General*
*Major General*
*Brigadier General*
*Colonel*
*…*

Military grades

*…*
*0-11*
*0-10*
*0-9*

Military facilities

*…*
*Headquarters*
*Unit*
*Command post*
*Observation Post*
*Installation*
*…*

General Service levels

*…*
*GS12*
*GS11*
*…*

Military Units

*…*
*Major Air Command*
*Air Command*
*Air Force*
*Air Division*
*Wing*
*Group*
*Squadron*
*Flight*
*…*

Military Unit Knowledge

*Operations support departments*
*Field operations*
*Typical commander rank*
*Typical number of personnel*
*Typical Facility type*
*…*

Links

*Ranks and military pay grades*
*Pay grades and incomes*
*GS levels and military grades*
*Ranks of different armed services*
*Ranks of different countries*
*…*

Chinese Air Forces

*Command Structure*
*Military Regions*
*…*

**Figure 3 The IAA-Cyc Project leveraged the Cyc KB capabilities to the benefit of IAA and its analyst users**

# 6 Summary of Accomplishments

## 6.1 IAA-Cyc Information Extraction (IE) Software Development

In the area of software development for information extraction, project accomplishments included design and development of the following:

- ***Framework:*** A "plug-in, plug-out" framework for the system that serves as the processing framework and testbed for the information extraction components.
- ***Lexicon and database:*** Software components such as the lexicon and the database for extracted information.
- ***Improved entity identification:*** Software to improve entity identification in text documents/messages.
- ***Coreference resolution and normalization:*** Software to perform coreference resolution and normalization of certain types of references (personal pronouns, proper names, and limited forms of descriptions).
- ***Attribute extraction:*** Software to automatically extract information concerning attributes about persons and organizations involving positions, units, ranks, postings and facilities.
- ***Relationship extraction:*** Software to automatically extract information concerning relationships among persons and organizations; relationships of interest include economic, familial, political, organizational, and religious relationships.
    - ***Noun phrase analysis:*** Software to perform noun phrase analysis to extract the above mentioned attributes and relationships.
    - ***Clause analysis:*** Software to perform analysis of clauses that express directly the above mentioned attributes and relationships.
- ***Meta-data extraction:*** Software that performs clause analysis to identify and normalize a date/time to associate with an extracted attribution or relationship, if possible.
- ***Information inference:*** Software that infers information that is not expressed explicitly in the text of a document; the information is inferred from extracted and known information.

For the previous IAA-Cyc 1 effort, all these components were primarily implemented within the Cyc KB, along with an associated C program that performed some querying and processing.

For the current IAA-Cyc 2 effort however, most of the components were re-implemented so as to be separate from the Cyc KB. The performance of the previous version of the IAA-Cyc software was not adequate to meet user requirements, especially in terms of performance speed.

The IAA-Cyc 2 software framework is implemented in C++ on the Win32 platform. The system includes certain OTS components such as the Text Zoner developed by Cymfony, Inc., the BBN IdentiFinder™ for named entity identification, and the Oracle DBMS for information storage and management. In the current IAA-Cyc 2 system implementation, the system is distributed across platforms, with the majority of the system residing on a Win32 platform. However, three components reside on Sun Unix-based machines. These latter three components are the Text Zoner, IdentiFinder, and the Oracle DBMS. A Unix Socket Server (USS) was developed to run on the Unix platform to provide the Win32-based components with access to the Unix-based applications required for information extraction (i.e., IdentiFinder and the Text Zoner).

## 6.2 Cyc KB Ontological Engineering

In addition to refinements and enhancements to the accomplishments of the Phase 1 project, the Cyc KB ontological engineering work on this IAA-Cyc 2 project primarily focused on an attempt to exploit the capabilities of the Cyc KB to reason about extracted information and to check the degree to which new extracted information conforms with or adheres to expectations or predefined patterns or "profiles". Our premise is that both conforming information and non-conforming information is valuable. It is important to be able to detect when a critical profile is matched, for example, when the profile is that of a terrorist. It is also important to be able to detect when anomalous conditions of interest occur, such as when an individual "rises through the ranks" of an organization much more quickly than would ordinarily be expected.

A high degree of utility would also attach to a system that could analyze a corpus of extracted data for anomalies. These anomalies might reflect flaws in either data collection methods or in the extraction process itself.

Speaking generally, the same desideratum underlies both objectives: namely, the ability to compare the information encoded in a structured database to general expectations in such a way as to detect divergence between expectations and represented facts. This ability is what we have endeavored to implement, in an initial, scalable form, in Cyc in the IAA-Cyc 2 project.

Accomplishments of the Phase 1 project that were refined include the ontological engineering (OE) of knowledge under the following general headings:

- Military Positions
- Anticipated and actual tenures (in ranks and in positions)
- Rank comparatives
- Faceting functions for military organizations
- Organizational facilities and postings
- Rank-to-position mappings
- Command structure of the Chinese PLA and PLAAF

CycL specifications of the internal command hierarchy and military force structure of the PLA and PLAAF deserve mention as presenting special technical considerations. Specifically, our source documents contained distinct descriptions of the PLA command structure at five distinct phases or levels of development: early history, 1947-1954, 1954-1970, 1970-1985, and 1985-present. Assertions that were true in one time frame were not necessarily true in any of the others.

Our solution to this problem was to sequester period-specific assertions into temporally indexed Cyc microtheories. These were specialized microtheories of a general ChineseMilitaryForceStructureMt microtheory whose assertions were presumed to hold throughout the ‘early history through 2001’ time frame. Although this solution was acceptable for the purposes of the project, it proved possible to implement only because knowledgeable members of our development team could make fairly hard-and-fast distinctions between assertions that held true for the PLA/PLAAF command structures generally (throughout all the time periods referenced) and assertions that held true in exactly one of the specified time periods. If we had to deal with ‘intermediate’ assertions that covered proper, non-singleton subsets of the set of time frames (e.g., 1947-1985), then it would have been necessary to reify a more complex partial order of microtheories.

# 7 Technical Approach and Accomplishments

## 7.1 IAA-Cyc Information Extraction Software Development

This section describes the approach and accomplishments of this project in the area of information extraction (IE) software development.

### 7.1.1 Development of IAA-Cyc Framework and Systems Engineering

A "plug-in, plug-out" system framework was developed that serves as the processing framework and testbed for the information extraction components. This framework is similar to the framework used for the IAA system. This type of framework supports modularization and facilitates development of the system. It provides for "plugging in" new technology as it becomes available, thus enabling the Government to leverage new technology developed as a part of other projects. This modularization also enables the most appropriate technology to be used for each step of document processing for information extraction.

#### 7.1.1.1 Speed up of IAA-Cyc 2 System

As part of this project, effort was applied to reducing the time it takes to run a document through the IAA-Cyc 2 system. The prototype resulting from the IAA-Cyc 1 effort was too slow to meet user requirements. The system framework and the majority of the components were re-implemented in C++, with selected components interacting with the Cyc KB for inference.

The slowest aspect of the IAA-Cyc system capabilities involved looking up terms in the database lexicon in order to determine their semantic types. To address this issue, we designed and implemented an approach to read the database lexicon tables into memory-resident "map" and "multimap" Standard Template Library (STL) data objects. These STL objects, which function as lookup lists based upon an index or key, were implemented to achieve fast search performance. These STL objects are now used to lookup terms in the IAA-Cyc 2 lexicon. Further speedup was accomplished by eliminating the need to load the STL maps afresh for each document and incorporating multi-threaded processes where necessary.

The speedup due to these changes was considerable. The IAA-Cyc 1 system processed documents at a rate of 1 ½ minutes per sentence. The new IAA-Cyc 2 system processes documents at a rate of approximately 4 seconds per document. The documents are typically comprised of 30 – 50 sentences.

### 7.1.2 Text Zoning

The first stage of processing by the IAA-Cyc 2 prototype consists of Text Zoning. The Text Zoner developed by Cymfony Inc. and used in the IAA system was incorporated into the IAA-Cyc 2 prototype. The Text Zoner performs the automated analysis of documents/messages to recognize and identify the parts of the documents/messages. These parts include:

- tagged information fields, such as subject, source, distribution list, date, country, etc. (tagged with a label such as "SUBJ:", "SOURCE:", etc.);
- the free text prose portions;
- structured parts such as tables, lists, etc.;
- separators, footnotes, etc.; and
- text that is extraneous to the actual content of a document, such as page breaks, page headers, page footers, etc.

.

### 7.1.3 Lexicon Development

The IAA-Cyc 2 lexicon development work focused on building a lexicon to enable and support the information extraction processes, including entity identification, coreference, attribution, and relationship extraction. The following subsections discuss this effort.

#### 7.1.3.1 Cross-Domain Lexicon Development

To maintain domain portability and domain independence to the greatest extent possible in the developed software system, the lexicon development effort focused on the incorporation of terms that are useable across many domains (e.g., ranks are general across many militaries, job positions such as "director" apply to many occupations, and nationalities pertain to many domains).

As part of this lexicon development effort, special emphasis was placed on developing the portion of the lexicon comprised of terms that express the attributes of *nationality*, *job position*, and *military rank*. The lexicon was developed so that the extraction processes could use the lexicon to identify entities, their relevant attributes, and extract the corresponding attributions and relationships.

#### 7.1.3.2 Domain-Specific Lexicon Development

In more domain-specific areas, emphasis was placed on entering terms that pertain to information that is highly useful and/or stable. For example, the set of names for the military ranks in a given country do not typically change much over time and is an example of stable terminology.

Terms that were entered into the lexicon included the military ranks of China and the specialized organizations of the Chinese military.

#### 7.1.3.3 Interfacing with the Cyc KB

Whenever the terms added to the term lexicon had corresponding constants in the Cyc KB, the corresponding KB constant was added to the term lexicon. The KB constants were used when inferring information from the KB. That is, when it was desirable to use the Cyc KB to infer information about the items of interest, certain attribution facts would be asserted to the KB using the KB constants and these facts used to trigger inferences in the KB.

### 7.1.4 Entity Identification

#### 7.1.4.1 Identification of Entities and References

Information Identification is the recognition of text segments comprising expressions for items such as entities, entity attributes, relationships, and simple events. This section is concerned only with the identification of expression referring to entities. These expressions can be categorized as follows:

- Names
    - Multi-token names: “Military Region”, “People’s Liberation Army”
    - Prepositional compounds: “Secretary of State”, “Commander of the PLAAF”
- Pronouns (with person, number and gender attributes)
- Descriptions
    - Definite: “the Unit Commander”
    - Indefinite: “a responsible officer”
- Lists
    - Qualification lists (subparts)
      Organizations: “Political Department, PLAAF Headquarters”
      Locations: “Paris, Texas”
      Temporal: “May, 1999”
    - Conjunctions and uniform lists
      “Korea, Japan”, “Clinton, Barak, Arafat”, “Clinton and Barak” etc.
- Appositives
    - Comma delimited: “PLAAF Commander, General Yu Zhenwu”
    - Not comma delimited: “PLAAF Commander General Yu Zhenwu”
- Parentheticals
    - Acronyms: “Military Region (MR)”

References are extracted from:

- Identified noun groups (Partial Parsing)
    - Noun groups are broken by verbs, conjunctions, prepositions and punctuation

- Identified named entities (IdentiFinder)
    - Persons, organizations, locations, times from IdentiFinder

- Recognized proper names (Lexicon Lookup)
    - Looked up in the Lexicon

- Interpretation of descriptions (IE)
    - Determination of how a word or phrase modifies another
    - Determination of how a word or phrase narrows or qualifies the meaning of an entity

References serve as temporary constants:

- For making assertions concerning meaning
- For making assertions concerning coreference

Example Input and Results:

> Input: “MGEN XU CHENGDONG, DIRECTOR, POLITICAL DEPARTMENT (PD), PLAAF HEADQUARTERS."

Results of selected processing stages:

- *Tokenizer:* “MGEN” “XU” “CHENGDONG” “,” “DIRECTOR”, “POLITICAL” “DEPARTMENT” “(“ “PD” “)” “,” “PLAAF” …
- *Entity identification:* “MGEN XU CHENGDONG”
- *Partial parser:* “MGEN XU CHENGDONG” “DIRECTOR” “POLITICAL DEPARTMENT (PD)” “PLAAF HEADQUARTERS”
- *Recognized proper names:* “POLITICAL DEPARTMENT” and “PLAAF”
- *Recognized common nouns:* “DEPARTMENT” and “HEADQUARTERS”
- *Recognized qualification relations:* “PLAAF” qualifies “HEADQUARTERS” “POLITICAL DEPARTMENT” qualifies “PLAAF HEADQUARTERS”

#### 7.1.4.2 Handling Results from Multiple Entity Identification Components

The IAA-Cyc 2 software system includes and uses multiple entity identification components. The purpose of having multiple entity identification components is to improve the recall and precision of the entity identification step over the performance that would be provided by just one entity identification approach. No single technology solves the entity identification problem, and the individual technical approaches have their strengths and weaknesses. For example, the statistical-based approach of IdentiFinder has the advantage that it can detect entity names that it has never encountered or seen before. Of course, it will not detect 100% of entity names, and its performance may suffer on text that is not similar to the type of text on which it was trained. So, complementary technology has been included in the IAA-Cyc system, namely a Lexicon Lookup component and natural language processing (NLP) components.

The entity identification components included and used by the IAA-Cyc 2 system are listed below. The list reflects the preferential ordering of results for entity names:

1. Lexicon Lookup results.
2. IdentiFinder results.
3. NLP Component results.

The Lexicon Lookup component was implemented to ensure that information about known entities of interest is not missed. In areas in which lists of names are available for many of the entities of interest, these lists have been loaded into the IAA-Cyc lexicon and are available to the Lexicon Lookup component for lookup and match during extraction processing.

The NLP component uses the results of part-of-speech tagging and partial parsing, and employs semantic categorization as part of its analysis.

As part of this project, we addressed the issue of handling and resolving entity identification results generated by the multiple entity identification components within the system. Algorithms were developed and incorporated into a component that detects and resolves conflicts among the entity identification results returned from the multiple components, namely the IdentiFinder component, the Lexicon Lookup component, and the NLP component.

The implemented algorithms deal with conflicts arising from the assignment of parts-of-speech, noun groups, and verb groups to the text. These assignments determine which text groups are looked up to identify entities. For example if the word, “SHANGHAI” is assigned a verb part of speech, it would not be looked up in the lexicon to determine if it is an entity. Also, if IdentiFinder recognized this word as a LOCATION, this identification would be ignored.

In order to address these problems, two recent changes were made to the algorithm for identifying entities. First, adjacent groups are now joined together and searched for entity term matches in the lexicon when partial matching within a single group indicates that a longer match might be found within the appended groups. For example, if the groups "GUANGZHOU" and "MILITARY REGION" are in succession, they would be joined together and a search performed to determine whether "GUANGZHOU MILITARY REGION" is in the lexicon. Second, if no lexicon entity matches are found, then the results of IdentiFinder are used to correct mis-assignments of noun group and verb group boundaries.

The IAA-Cyc 2 conflict resolution approach still needs additional work and will undergo continued development in the next project phase.

### 7.1.5 Acquiring Acronyms and Their Definitions

We developed an algorithm for recognizing and acquiring acronyms from the text documents in which they appear. The algorithm has been developed and implemented to recognize when an acronym appears in parentheses following the proper noun or name that defines it and for which the acronym is then an alias. When a new acronym is detected, it is added to the in-memory acronym list for the document. An example would be “Political Department (PD)”. For this example, the algorithm recognizes "PD" as an acronym for “Political Department” and adds the acronym to the in-memory acronym list. The new acronyms are then used for within-document coreference resolution.

The acronym algorithm handles situations in which the noun group preceding an acronym includes more text than should be associated with the acronym. For example, in the noun group “THE GUANGZHOU MILITARY REGION (MR)” only the text “MILITARY REGION” should be associated with the acronym “MR”.

The acronym algorithm needs refinement so as to handle cases when the letters of the acronym do not simply match the first letters of all the words of the text that should be associated with the acronym (e.g., “The Southeast Asian Treaty Organization (SEATO)”).

An area of future work is the manner in which acronyms should be added to the persistent lexicon. Since an acronym can frequently have different meanings depending on the context (e.g., "PD" could represent "Political Department" or "Police Department", among others), the addition of acronyms could be added to domain contexts in persistent storage, subject to user approval.

### 7.1.6 Conjunctive Reference Identification

As part of the IAA-Cyc 2 project, software was implemented that groups two entities together into a joint entity. This new joint entity can then serve as an antecedent to a plural pronoun. For example, the joint reference "Britain and Zimbabwe" can corefer to "they" or "we" in the text of a document.

### 7.1.7 Discourse Context

As part of the IAA-Cyc project, we designed and developed a technical approach to maintain a discourse context model and use it for coreference resolution and attribution. A discourse model component was developed to identify and track information comprising the discourse context of a document. The discourse context of a document may be used to help perform operations such as the identification of the antecedents of pronouns, attributions in the text, and the date/time at which or during which an attribution might hold.

The discourse context model consists of a conceptual object that includes a representation of selected elements that comprise the focus of a document's discourse. Information is gathered to create the discourse context data structures associated with the entities identified in the text. This includes information in the document that expresses:

- who - the person that is the focus of the discourse,
- when - the time frame of the discourse, and
- what - the object or event that the discourse concerns.

The “who” associated with a pronominal entity is a likely candidate for the antecedent of a pronoun. The “what” associated with entities and attributes is used to determine attributions expressed by clauses (e.g., if the “what” associated with a person entity and a job position attribute, such as “director”, is a job position verb, such as “promoted”, then a possible attribution is identified).

Algorithms (methods) were designed and implemented as part of the discourse context model. Included among these are algorithms to maintain the currency of the discourse context during the processing of a document, and algorithms to retrieve information from the discourse model and return it when requested.

The discourse context model is used for coreference resolution. The discourse context data structure is used to maintain a focal “who” that indicates the person that is being primarily referred to (the “focus” of the discourse) within the current discourse unit (e.g., the current clause or sentence). The discourse model includes capabilities for tracking and representing the current male, female, and plural "who". The “who” is then used to determine the person or persons to whom a personal pronoun refers (e.g., an instance of “he” would be assigned, as a referent, the normal form of the current male discourse “who”).

The discourse context data structure is also used to assign meta-data to extracted attributions. For example, the discourse data structure maintains a focal “when” that indicates the time frame that is the current focus of the discourse. This “when” is then used when assigning date/time meta-information to the extracted attributions (see discussion in subsequent section).

#### 7.1.7.1 Identification of Speech Contexts

Two special types of contexts involving reported speech were identified and incorporated into the discourse context model in order to support the resolution of pronoun coreferences.

The first context is for *directly quoted speech* such as the sentence "the General said, 'The 8th Army Unit will be deployed next week' ".

The second context involves a *speaker bracketing convention* used when reporting interviews within a document. This convention uses bracketed text to identify the speaker. Examples include: "[Reporter] When will you travel to Serbia?" and "[Medovic] I supported Milosevic."

These contexts are identified by the IAA-Cyc software so that a representation of the Current Speaker may be maintained by the discourse context model software to resolve personal pronoun references such as "I", "me", "we", etc.

### 7.1.8 Nominal Reference: Merging of Mention and Discourse Reference Information

As part of IAA-Cyc 2, the use of discourse references and mentions as separate data objects was eliminated. The two types of items were represented separately in the technical approach of IAA-Cyc 1.

These two conceptual objects were merged into a single conceptual object type, which we call a *nominal reference*. These nominal references include names, pronouns, and descriptions. The differences between mentions and discourse references may still be identified in that a discourse reference is essentially a nominal reference that is not contained within another nominal reference.

The identification of compound nominal references (formerly discourse references) was also changed. Previously, these references were identified by grouping noun groups until a "stop" condition was met. Stop conditions included verbs, certain punctuation and prepositions etc., that indicate when a reference ends and another reference, verb group, or other sentence element (e.g., adverb or adjective group) begins. This approach to identifying references, with its "greedy" nature in constructing long references, greatly reduced the number of references that needed to be considered for coreference. However, there were cases in which coreferences were missed because the antecedents were hidden within these long references (e.g., due to "over attachment" of prepositional phrases).

A more conservative approach is now taken. Nominal references are now grouped together only when they may be semantically associated. For example, "Prime Minister of Bulgaria" is identified as a compound reference (of the references "Prime Minister" and "Bulgaria") because a semantic association is identified between the two (in the example, an association is determined when a reference is identified with the Position semantic class followed by "of" and a Location reference).

### 7.1.9 Coreference Capability

A coreference capability was developed for the IAA-Cyc prototype to address the problem of identifying expressions in a text document that refer to the same entity.

The processing for coreference consists of the following capabilities:

- Preprocessing that identifies discourse references, mentions, and features in the text which are used in the determination of coreference.
- Identification of name coreferences (i.e., alias detection).
- Identification of pronominal coreferences.
- Identification of description coreferences.

#### 7.1.9.1 Preprocessing

Preprocessing is responsible for identifying nominal references such as noun groups that are unambiguous noun phrases identified by the partial parser (e.g., "Bulgaria", "the military facility") and entity names such as for persons, organizations, locations, and dates/times.

The coreference component, which is run after the preprocessing step, depends on certain features of identified references which are assigned during preprocessing. The two most important features are:

- the identified *grammatical function* of the reference within an identified clause such as subject, predicate, direct object, indirect object, temporal adjunct, locative adjunct, and unclassified adjunct, and
- the *identified role* of the reference in the representation of the discourse context focus (i.e., the data structure that specifies the dynamic Who, What, Where, and When that is involved in each sentence of the discourse).

Additional features that are used include number (e.g., singular, plural), gender, and animacy.

#### 7.1.9.2 Coreference and Normalization Processing

Coreference and normalization capabilities were developed as part of the IAA-Cyc 2 effort. These processes consist of the following steps:

1. Loading into C++ Objects noun group and verb group information from the partial parser.
2. Identifying the features of noun groups that are relevant for coreference:
   - Lexicon lookup to determine the semantic types and normal forms of words and compound words (e.g., "Prime Minister").
   - Determining features of number, person, animacy, pronoun case (i.e., subjective, possessive, objective), pronoun class (i.e., personal, relative, interrogative, definite, indefinite), definiteness (i.e., definite, indefinite) from parts-of-speech and pronoun text.
3. Identifying adjective, adverb, conjunction, preposition, and punctuation groups from the part-of-speech information from the partial parser.
4. Handling of the following special cases:
   - Conjunctive references such as "the Foreign Ministers of France and Canada" or "the French and Canadian Foreign Ministers" that need to be transformed into separate references. For these examples, the resulting separate references would be "the Foreign Minister of France" and "the Foreign Minister of Canada" for the first example phrase, "the French Foreign Minister" and "the Canadian Foreign Minister" for the second example phrase.
   - Appositives such as "Bulgarian Prime Minister Zhan Videnov" that need to be transformed into separate references, namely "Bulgarian Prime Minister" and "Zhan Videnov" for this example phrase.
   - Locative expressions such as "Paris, Texas" (two references, namely "Paris" and "Texas") that need to be transformed into a single reference ("Paris, Texas").
   - Metonymical references involving countries (e.g., "Yugoslavia sent Milosevic to the Hague today" where "Yugoslavia" is a reference to the Yugoslav government) that

need to be assigned their proper semantic types (e.g., Government Organization rather than Location).

5. Identifying Clause and Clause Constituent C++ Objects from the group information.
6. Identifying the discourse context (i.e., the who, what, where, and when) of each sentence.

### 7.1.9.3 Name Coreference

Name coreference determines when identified names in the text refer to the same entity. The name coreference algorithm developed for the IAA-Cyc prototype checks for matches between one text reference and another. The matching criteria are based upon:

1. Aliases within the document that are indicated by parentheticals (e.g., "People's Liberation Army [PLA]").
2. Aliases stored in the database (e.g., "KFOR" as an alias for "Kosovo Force").
3. Normal forms derived according to syntactic features (e.g., "U.N. Security Council" for "the U.N. Security Council").
4. Normal forms stored in the database (e.g., "Kosovo Force" for "KFOR").
5. Variants of names (e.g., "Videnov" for "Zhan Videnov").
6. Text matching, when the semantic types assigned the text references are consistent ("Arpad Goncz" matches "Arpad Goncz").

### 7.1.9.4 Pronoun Coreference

Pronoun coreference determines the most likely antecedent for a pronoun in the text. The implemented algorithm for pronoun coreference filters potential antecedents from the sentence containing the pronoun and from the preceding sentence. This filtering is based upon:

1. Grammatical constraints (e.g., a pronoun does not corefer with a clausal coargument; in the clause "The PLA Commander promoted him", the direct object "him" cannot corefer to the subject "The PLA Commander").
2. Mismatches of the features of number (e.g., "the commander" and "they"), gender ("the man" and "her"), and animacy ("the commander" and "it").
3. Mismatches in known semantic classes (e.g., "he" and "the U.N." where "the U.N." has been identified as an organization.

The potential antecedents that are not filtered according to these criteria are ranked according to salience. The numeric salience score is calculated based upon several features. These are:

1. Grammatical function (e.g., a clause subject is assigned a higher score than a clause direct object).
2. Semantic class (e.g., an identified person is assigned a higher score than a reference that has an unknown semantic class when evaluating potential antecedents for a singular personal pronoun).
3. Distance of the potential antecedent from the pronoun (the candidate references in the same sentence are scored higher than those in the preceding sentence).

4. Whether the reference is preceded by a preposition.

After salience scoring of the filtered potential antecedents for a pronoun, the highest scoring reference is selected as an antecedent. If all possible antecedents have been filtered out, then no antecedent is determined for the pronoun. The second and third best choices for an antecedent are saved so that they are available if the best choice subsequently is rejected according to some other criteria. This approach is based on the work of Lappin & Leass.

### 7.1.9.5 Description Coreference and Normalization of Positions

A capability was designed and implemented for identifying the antecedents of descriptions of positions (e.g., "HIS PRESENT POSITION", "HIS SHANGHAI POST"). A capability was also developed to normalize these descriptions based upon the identification of their antecedents.

The algorithm depends on the proper identification and normalization of :

- the antecedent of any genitive modifier (e.g., "HIS" or "GUANGZHOU MR'S")
- the temporal modifier (e.g., "PRESENT")
- the locative modifier (e.g., "SHANGHAI")
- the organization modifier (e.g., in "HIS 7TH AIR ARMY COMMAND")

The algorithm also depends on the identification of the following attributes of position references within the text:

- the person that holds the position.
- when the position was held.
- the physical location of the position.
- the organization within which the position exits.

After the antecedent of a position description is identified, the description is normalized and the surrounding context is used to determine attributions of the position. For example, in the description "HIS PRESENT POSITION IN CHENGDU", the physical location attribute of the antecedent position is updated. The surrounding context may also involve a restatement of the position, as in "HIS PRESENT POSITION OF PD DIRECTOR, PLAAF HEADQUARTERS". In this case, the restatement of the position is omitted from further analysis so that redundant attributions are not produced.

This capability was extended to determining coreferences for a greater number of description phrases within a document. The IAA-Cyc software was extended so that it detects coreference between two definite references that may contain adjective modifiers (e.g., "the hard neo-communist line" and "the hard line"; "the ruling party" and "the party"). Currently, the software assumes that definite references of these forms are coreferring when they occur within a 300-character window. In the future, this character window will be able to be adjusted via parameter input.

The IAA-Cyc software was further enhanced so that appositive descriptors (e.g., "Zhan Videnov, Bulgarian Prime Minister") are also detected as coreferring (e.g., the descriptor "Bulgarian Prime Minister" corefers with "Zhan Videnov").

**Example Inputs and Results:**

Input:
"MGEN XU CHENGDONG, DIRECTOR, POLITICAL DEPARTMENT (PD), PLAAF HEADQUARTERS. LITTLE IS KNOWN OF XU'S PAST. HE WAS FIRST NOTED IN PRESS REPORTS IN MARCH 1992 AS A RESPONSIBLE OFFICER IN THE GUANGZHOU MILITARY REGION (MR)."

Results:

- The parenthetical "PD" corefers with "POLITICAL DEPARTMENT".
- The named entity "XU" corefers with "XU CHENGDONG".
- The pronoun "HE" corefers with the named entity "XU".
- The parenthetical "MR" corefers with "MILITARY REGION".

### 7.1.9.6 Composition of Normal Forms for Discourse References

The use of semantic associations is also vital in the composition of compound noun groups (formerly discourse references). The following syntactic types are now identified and their normal forms determined based upon semantic associations:

1. Noun Series
(e.g., "DIRECTOR, POLITICAL DEPARTMENT (PD), PLAAF HEADQUARTERS" is assigned the within-document normal form "PLAAF HEADQUARTERS POLITICAL DEPARTMENT DIRECTOR")

2. Prepositional Series
(e.g., "THE PD DIRECTOR OF THE PLAAF HEADQUARTERS" is assigned the within-document normal form "PLAAF HEADQUARTERS POLITICAL DEPARTMENT DIRECTOR")

Normalization also takes into account the part of speech assigned to each term in a reference. In general, words in a reference, which are neither proper nor common nouns, are stripped off when determining normal forms (e.g., "A YOUNG COMMANDER OF AN AIR GROUP" is normalized as "AIR GROUP COMMANDER"). An exceptional case involves possessives (e.g., "GUANGZHOU MR'S 7TH AIR ARMY"). In these cases, the possessive proper nouns are used to split a reference into two separate normalized references (e.g., "GUANGZHOU MILITARY REGION" and "7TH AIR ARMY").

### 7.1.10 Handling Attributes of Persons, Organizations, and Positions

IAA-Cyc 2 program accomplishments include the extraction of certain types of attributes of persons and organizations by the software prototype. Attributes of Persons, Organizations, and Positions are determined from:

- The surrounding context of the text reference:
  - Modifiers of the reference.
  - Preceding/following references.
- Information from the clause:
  - Clauses are classified according to their main verb.
  - Currently, only main verbs that directly express attributes are handled.

The attributes of persons that are extracted:

- A person's name and aliases.
- A person's position and their position within an organization.
- A person's title and/or military rank.
- A person's age.

The attributes of organizations that are extracted:

- An organization's name and aliases.
- The positions within an organization.
- The location of an organization.
- The suborganizations of an organization.

The attributes of a position attributed to a person that are extracted:

- The time period during which the person has the position.

An object-oriented approach has been taken in this IAA-Cyc 2 project to the representation of persons, organizations, and other entities and their attributes. All attributes of an entity are collected into their respective C++ data objects (Person, Organization, and Position objects).

A means of effectively handling temporal aspects of attributes was developed. To access information according to the date/time that it holds true, the information is indexed using TimeSnapshot objects. A TimeSnapshot object represents a period of time, and items of interest that are true during part of the TimeSnapshop time period (e.g., an entity having a certain attribute) are associated with the particular TimeSnapshop. A TimeSnapshot provides a means to gather relevant information that holds true during part or all of the period of interest. TimeSnapshots are used to compare and order attribute data based on temporal information attached to attributes. A TimeSnapshot object also has a time granularity assigned to the date/time (e.g., year, month or day) which gives the specificity of the date/time.

#### 7.1.10.1 Attributes and Attributions

Refinements were applied to the representation and storage of entity attributes in the IAA-Cyc database and the linking of those attributes to entities in the IBOK. A table entitled

"ATTRIBUTION_INSTANCE" was created that uniquely identifies particular instances of entities and their sets of related attributions. The attributions, in turn, identify a flexible and ordered chain of attributes about the entity. For example, an instance of the person "Xu Chengdong" has the following attribution: "Director" (an attribute) of the "Political Department" (an organization) of the "Headquarters" (a facility) of the "PLAFF" (an organization). The instance of "Xu Chengdong" will become a row in the "ATTRIBUTION_INSTANCE" table, and it will be linked to all of the attributions expressed in the "ATTRIBUTION" table. The instance will have metadata associated with it including locative and temporal information. Additional instances of "Xu Chengdong" may be created that describe other attributes of him at different locations or times. Each of these additional instances will be linked to another set of attributions in the "ATTRIBUTION" table.

### 7.1.11 Attribution Approach

This section discusses the design of the software capability for extracting and assigning attributes to entities.

The algorithm for determining attributions consists of three major steps:

- Lexicon lookup
- Attribution pattern lookup
- Database updates

Each of these steps is briefly discussed below.

#### 7.1.11.1 Lexicon Lookup

The lexicon is accessed for determining the semantic types of terms. It is also used for determining other semantic features useful for coreference identification, normal forms for the normalization process, and arguments (e.g., noun complements such as "of England" in "The King of England") for the identification of compound words and phrases. The identification of semantic types is necessary for recognizing attributions based upon type information (see next step).

The specification for a term lexicon entry is given below. The `TERM` field contains the actual word that comprises the lexeme. It is the term that the head tokens of identified noun groups are matched against. The `NORMAL_FORM` and `ROOT_FORM` are forms used in normalization. The `SEMANTIC_TYPE` field is a string that matches the name of an entity type in the database or knowledge base (e.g., `PERSON, ORGANIZATION, COUNTRY`), while the `SEMANTIC_ID` field is a pointer to an entity (e.g., a record of an entity table such as Person, Organization etc.). The `NUMBER`, `GENDER`, and `ANIMACY` fields are semantic features used in determining coreference (along with the `SEMANTIC_TYPE`). Finally, the `ARGUMENT` fields are used to recognize compound words.

```
TERM_LEXICON
 ------------------------------
 OBJECT_ID
 TERM
 NORMAL_FORM
 ROOT_FORM
 SEMANTIC_TYPE
 SEMANTIC_ID
 GENDER
 NUMBER
 ANIMACY
 ARGUMENT_TYPE
 ARGUMENT
 ARGUMENT_MARKER
```

#### 7.1.11.2 Attribution Rules

Determining when an attribute is an attribute of an entity (i.e., an attribution) is accomplished using a rule-based approach. The conditional antecedents of the rules are comprised of patterns, and the consequents are actions that generate attributions. The attribution process uses the information from a match of an attribution pattern condition (i.e., a specified attribute and attribute holder type) to produce information for insertion into the user's Interim Body of Knowledge (IBOK - see below).

The semantic types of the terms in a noun group, as determined through lexicon lookup or named entity identification processes (e.g., IdentiFinder or the Text Zoner), are used to create a pattern representation of the noun group (e.g., "`POSITION PERSON`" for "President George W. Bush"). This pattern is used to match against the attribution patterns comprising the antecedents of the rule set. The specification of an `ATTRIBUTION_PATTERN` is provided below, each of which has an a corresponding consequent output structure represented by an `ATTRIBUTION_LEXICON` structure (see specification below).

When there is a match, the `ATTRIBUTE`, `ATTRIBUTE_HOLDER`, and possibly `OUTPUT` fields of the `ATTRIBUTION_LEXICON` structure are used as a specification for producing an information structure for insertion into the IBOK or updating the IBOK with the attribution information (see next step). For example, the pattern "`POSITION PERSON`" would be associated with `ATTRIBUTE="POSITION"` (the target attribute of the search) and with `ATTRIBUTE_HOLDER="PERSON"` (outputs that do not match these table names and other output fields would be specified using the `OUTPUT` fields). The attribution identification process uses the text strings associated with these semantic types to produce information for the IBOK.

Our current implementation uses a database to hold the pattern-based rules described above. The storage and use of patterns in the database is an implementation decision which will be evaluated with respect to its speed and memory efficiency.

The ability of the attribution patterns in the database to express complex cases of attribution will also be evaluated. Many extraction systems use complex pattern specification languages to specify patterns. The intent of IAA-Cyc 2 is to initially use simple patterns and make use of context and semantic knowledge from the conformity checking capability to further "refine" identified attributions.

```
ATTRIBUTION_PATTERN
-------------------------------
OBJECT_ID
ATTRIBUTION_LEXICON_ID
PATTERN_STRING
```

```
ATTRIBUTION_LEXICON
-------------------------------
OBJECT_ID
ATTRIBUTE
ATTRIBUTE_HOLDER
OUTPUT_TABLE_1
OUTPUT_FIELD_1
OUTPUT_TABLE_2
OUTPUT_FIELD_2
```

#### 7.1.11.3 IBOK Update

Once an attribution is identified, the attribution information is stored in the IBOK tables in the database. Multiple attributions for a section of text are possible, and any conflicts in attributions will be resolved using the reasoning facilities of the conformity checking capability implemented in the Cyc KB.

Three database tables are updated with attributions. First, the ATTRIBUTION table (see specification below) which contains the basic information concerning the attribution: the attribute, attribute holder, document, and text offsets of the attribution.

Second, the ATTRIBUTE table (see specification below) associated with the type of attribution (e.g., JOB_POSITION_ATTRIBUTE table - see specification below) which contains information concerning when the attribute holds true of the attribute holder (TIME_HOLDS) and the previous and next values of the attribute for the attribute holder (e.g., PREVIOUS="GOVERNOR" and NEXT="EX-PRESIDENT" for "President Bill Clinton").

Finally, the ATTRIBUTE table often points to a table that provides information concerning the attribute, independent of any particular attribute holder. For example, the JOB_POSITION table row of "PRESIDENT" (that has the AFFILIATION attribute of "U.S. GOVERNMENT") will have TYPE="POLITICAL LEADER", PREVIOUS="PRESIDENTIAL CANDIDATE", and NEXT="EX-PRESIDENT" (in general, the previous and next fields will be used to specify job hierarchies; U.S. President is not a particularly good example of a position in a hierarchy).

```
ATTRIBUTION
-------------------------------
OBJECT_ID
ATTRIBUTE_TYPE
ATTRIBUTE
```

```
ATTRIBUTE_HOLDER_TYPE
ATTRIBUTE_HOLDER
EXPRESSION_TYPE
EXPRESSION
DOCUMENT
BEGIN_TEXT_OFFSET
END_TEXT_OFFSET

JOB_POSITION_ATTRIBUTE
-------------------------------
OBJECT_ID
JOB_POSITION_ID
TIME_HOLDS
PREVIOUS
 NEXT

JOB_POSITION
-------------------------------
OBJECT_ID
NAME
TYPE
PREVIOUS
NEXT
```

### 7.1.12 Attribution Capabilities

The IAA-Cyc attribution rules can be grouped into three main categories based on the type of text segment they are designed to handle. The three categories are:

- *Group:* Rules to analyze noun groups for attribution expressions. For example, the noun group "Prime Minister Tony Blair" would result in the system's recognizing and extracting "Prime Minister" as the Position of "Tony Blair".
- *Phrase:* Rules to analyze phrases. For example, the phrase "the President of the U.S." would result in the system's recognizing and extracting "President" as a Position of the United States.
- *Clause:* Rules to analyze clauses for attribution expressions. For example, the clause "Hillary Clinton is a Senator for New York" would result in the system's extracting "Senator" as the Position attribute of "Hillary Clinton".

In the case of *clauses*, the system recognizes verbs of the following for extracting attributions and associating date/time meta-information with attributions:

- *Reporting Verbs*: express the statement of information or the reporting of information (e.g., “said”, “reported”, “identified”, etc.)
- *Existence Verbs*: express the existence of entities or states, conditions, or attributes of entities (e.g., “is”, “was”, “became”, etc.)
- *Job Position Verbs*: express actions particular to the attribution of job positions (e.g., “promoted”, “demoted”, etc.)

Members of these verb classes are recognized by looking up the verbs (encountered in the text document) in the lexicon. More specifically, this currently means querying the VERB_LEXICON table in the IAA-Cyc Database with the main verbs within identified verb groups.

**Example Inputs and Results:**

Input:
"MGEN XU CHENGDONG, DIRECTOR, POLITICAL DEPARTMENT (PD),
PLAAF HEADQUARTERS. LITTLE IS KNOWN OF XU'S PAST. HE WAS
FIRST NOTED IN PRESS REPORTS IN MARCH 1992 AS A
RESPONSIBLE OFFICER IN THE GUANGZHOU MILITARY REGION
(MR)."

Results:

- The present rank of Xu Chengdong is Major General.
- The present position of Xu Chengdong is PLAAF Headquarters Political Department Director.
- The position in March 1992 of Xu Chengdong was Guangzhou Military Region officer.

Input:
"HE WAS NEXT SEEN IN GUANGXI IN 1984, PROBABLY ATTACHED TO THE
7TH AIR ARMY THERE. IN FEB 1991 HE SURFACED IN
SHANGHAI AS COMMANDER OF THE AIR FORCE UNIT THERE,
THE SHANGHAI COMMAND POST, WHICH IS A CORPS-LEVEL OR
MGEN LEVEL UNIT. IN MARCH 1992 HE WAS IDENTIFIED WITH
THE RANK OF MGEN."

Results:

- "He" refers to Huang Hengmei (through coreference processing).
- The location of Huang Hengmei in 1984 was Guangxi.
- The position of Huang Hengmei in February 1991 was Shanghai Command Post Commander.
- The location of Huang Hengmei in February 1991 was Shanghai.
- The rank of Huang Hengmei in March 1992 was Major General.

### 7.1.13 Relationship Extraction

Development of relationship extraction capabilities focused on relationships in the high-level categories of economic, political, family, organizational, and religious relationships. The relationships targeted for extraction include superior-subordinate, employer-employee, father-son, ally, and opponent relationships, among others.

Relationship extraction algorithms were developed and incorporated into the IAA-Cyc system for the targeted types of relationships. These algorithms involve the application of pattern-matching technology to detect relationship indicator words and the arguments to a relationship. As in the case of attributes, the IAA-Cyc relationship rules can be categorized into three main categories based on the type of text segment they are designed to handle. The three categories are:

- *Group:* Rules to analyze noun groups for relationship expressions. For example, the noun group "British Prime Minister Tony Blair" would result in the system's recognizing and extracting an affiliation relationship between Tony Blair and Great Britain.
- *Phrase:* Rules to analyze phrases. For example, the phrase "Milosevic's business partner Ivan Stambolic" would result in the system's recognizing and extracting an economic relationship between Milosevic and Stambolic. As another example, the phrase "Mira Markovic, the wife of Milosevic" would result in the system's recognizing and extracting a familial relationship between Mira Markovic and Milosevic.
- *Clause:* Rules to analyze clauses for attribution expressions. For example, the clause "Hillary Clinton is a Senator for New York" would result in the system's recognizing and extracting an affiliation relationship between Hillary Clinton and New York.

### 7.1.14 Inference of Information

#### 7.1.14.1 Inference of Information from Known and Extracted Information

Research and development was performed in the area of inferring information from known and extracted information. As part of this effort, initial information inference capabilities were developed for certain types of attributes and relationships.

For example, a capability was developed to infer a person's relationship to an organization and their probable physical location during the time interval in which they held a certain position. Both inferences depend upon the attribution of a position to an individual. Currently, the inference rules used are:

- If a person holds a certain position and that position is within an organization, then the person has a job affiliation with that organization.
- If a person holds a certain position and that position is located at a certain physical location, then the person is located at that location.

Attributions between a job position and an organization within a single sentence are derived when the following conditions are met:

- A person mentioned in the sentence has been associated with an organization.
- The person has also been associated with a job position.

Effort was also applied to developing an approach for inferring work relationships between two persons. However, the design and implementation for this approach has not yet been completed. These inferences will determine when one person worked for, supervised, or worked with another person.

This work is on-going.

#### 7.1.14.2 Inference of Meta-Information

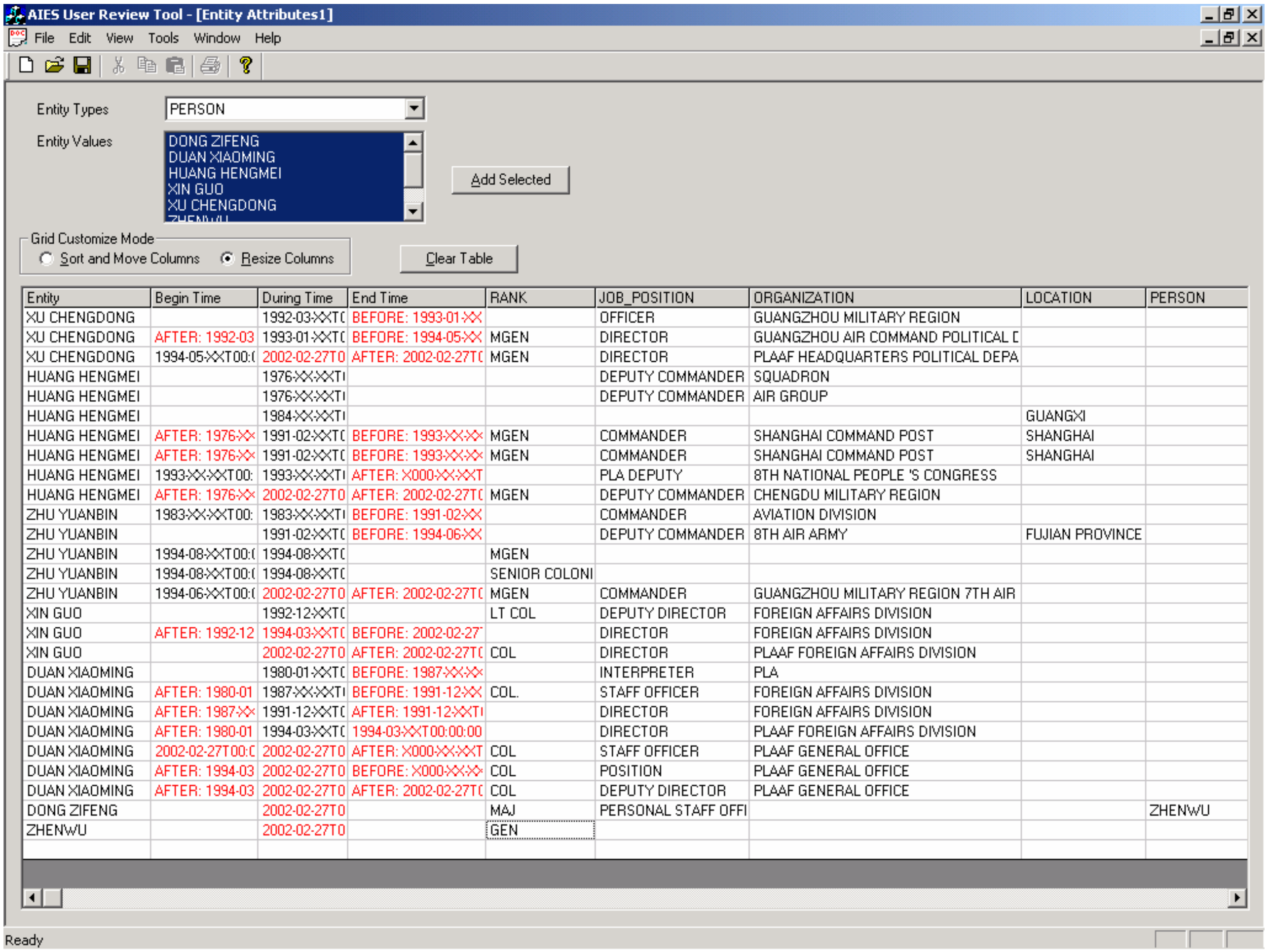
Research and development was performed in the area of inferring meta-information from known and extracted information, especially temporal meta-information. As part of this effort, we developed a prototype capability to infer when a person began and ended any type of position. The inference of this information makes use of the following assumptions:

- If a person was reported to hold a position during a certain time interval or to have ended a position at a certain time, then another position reported to hold within a later time frame *must have begun after the end* of first position's reported time frame.
- Conversely, a position with an earlier time frame is inferred *to have ended before* a position with a later time frame.

These inferences are relaxed or inhibited when it is determined from the Cyc KB that the two particular positions may be held concurrently. This work in on-going.

### 7.1.15 GUI for Attribution Results Display

A graphical user interface (GUI) was developed for the display of entity attributes in tabular report form. An example of the Entity Attributes tabular report display is shown in the figure below. This tabular report displays extracted entities and their extracted attributes. These information items are retrieved from the analyst's Interim Body of Knowledge (IBOK) portion of the IAA-Cyc database. Via the user interface, the analyst has the capabilities to restrict entities by type, select single or multiple entities, and display the aggregated attributes of all selected entities in an ordered and sorted table. The analyst may reorder/resize columns and resort the data with respect to any column. The ability to display multiple entities simultaneously facilitates comparing and contrasting information on the entities. This GUI for displaying attribution results was developed using Microsoft Access. This GUI is a preliminary version. The eventual goal is to enable the user to view and evaluate the results using a more comprehensive graphical user interface. The next section provides a more detailed explanation of the extraction results displayed in the figure.



Entity Types: PERSON

Entity Values:
DONG ZIFENG
DUAN XIAOMING
HUANG HENGMEI
XIN GUO
XU CHENGDONG
ZHENWU

Grid Customize Mode: ○ Sort and Move Columns ● Resize Columns

| Entity | Begin Time | During Time | End Time | RANK | JOB_POSITION | ORGANIZATION | LOCATION | PERSON |
|---|---|---|---|---|---|---|---|---|
| XU CHENGDONG | | 1992-03-XXT( | BEFORE: 1993-01-XX | | OFFICER | GUANGZHOU MILITARY REGION | | |
| XU CHENGDONG | AFTER: 1992-03 | 1993-01-XXT( | BEFORE: 1994-05-XX | MGEN | DIRECTOR | GUANGZHOU AIR COMMAND POLITICAL [ | | |
| XU CHENGDONG | 1994-05-XXT00:( | 2002-02-27T0 | AFTER: 2002-02-27T( | MGEN | DIRECTOR | PLAAF HEADQUARTERS POLITICAL DEPA | | |
| HUANG HENGMEI | | 1976-XX-XXT( | | | DEPUTY COMMANDER | SQUADRON | | |
| HUANG HENGMEI | | 1976-XX-XXT( | | | DEPUTY COMMANDER | AIR GROUP | | |
| HUANG HENGMEI | | 1984-XX-XXT( | | | | | GUANGXI | |
| HUANG HENGMEI | AFTER: 1976-XX | 1991-02-XXT( | BEFORE: 1993-XX-XX | MGEN | COMMANDER | SHANGHAI COMMAND POST | SHANGHAI | |
| HUANG HENGMEI | AFTER: 1976-XX | 1991-02-XXT( | BEFORE: 1993-XX-XX | MGEN | COMMANDER | SHANGHAI COMMAND POST | SHANGHAI | |
| HUANG HENGMEI | 1993-XX-XXT00: | 1993-XX-XXT( | AFTER: X000-XX-XXT | | PLA DEPUTY | 8TH NATIONAL PEOPLE 'S CONGRESS | | |
| HUANG HENGMEI | AFTER: 1976-XX | 2002-02-27T0 | AFTER: 2002-02-27T( | MGEN | DEPUTY COMMANDER | CHENGDU MILITARY REGION | | |
| ZHU YUANBIN | 1983-XX-XXT00: | 1983-XX-XXT( | BEFORE: 1991-02-XX | | COMMANDER | AVIATION DIVISION | | |
| ZHU YUANBIN | | 1991-02-XXT( | BEFORE: 1994-06-XX | | DEPUTY COMMANDER | 8TH AIR ARMY | FUJIAN PROVINCE | |
| ZHU YUANBIN | 1994-08-XXT00:( | 1994-08-XXT( | | MGEN | | | | |
| ZHU YUANBIN | 1994-08-XXT00:( | 1994-08-XXT( | | SENIOR COLONI | | | | |
| ZHU YUANBIN | 1994-06-XXT00:( | 2002-02-27T0 | AFTER: 2002-02-27T( | MGEN | COMMANDER | GUANGZHOU MILITARY REGION 7TH AIR | | |
| XIN GUO | | 1992-12-XXT( | | LT COL | DEPUTY DIRECTOR | FOREIGN AFFAIRS DIVISION | | |
| XIN GUO | AFTER: 1992-12 | 1994-03-XXT( | BEFORE: 2002-02-27 | | DIRECTOR | FOREIGN AFFAIRS DIVISION | | |
| XIN GUO | | 2002-02-27T0 | AFTER: 2002-02-27T( | COL | DIRECTOR | PLAAF FOREIGN AFFAIRS DIVISION | | |
| DUAN XIAOMING | | 1980-01-XXT( | BEFORE: 1987-XX-XX | | INTERPRETER | PLA | | |
| DUAN XIAOMING | AFTER: 1980-01 | 1987-XX-XXT( | BEFORE: 1991-12-XX | COL. | STAFF OFFICER | FOREIGN AFFAIRS DIVISION | | |
| DUAN XIAOMING | AFTER: 1987-XX | 1991-12-XXT( | AFTER: 1991-12-XXT( | | DIRECTOR | FOREIGN AFFAIRS DIVISION | | |
| DUAN XIAOMING | AFTER: 1980-01 | 1994-03-XXT( | 1994-03-XXT00:00:00 | | DIRECTOR | PLAAF FOREIGN AFFAIRS DIVISION | | |
| DUAN XIAOMING | 2002-02-27T00:( | 2002-02-27T0 | AFTER: X000-XX-XXT | COL | STAFF OFFICER | PLAAF GENERAL OFFICE | | |
| DUAN XIAOMING | AFTER: 1994-03 | 2002-02-27T0 | BEFORE: X000-XX-XX | COL | POSITION | PLAAF GENERAL OFFICE | | |
| DUAN XIAOMING | AFTER: 1994-03 | 2002-02-27T0 | AFTER: 2002-02-27T( | COL | DEPUTY DIRECTOR | PLAAF GENERAL OFFICE | | |
| DONG ZIFENG | | 2002-02-27T0 | | MAJ | PERSONAL STAFF OFFI | | | ZHENWU |
| ZHENWU | | 2002-02-27T0 | | GEN | | | | |

**Figure 4 The IAA-Cyc Entity Attributes Table Report displays the entity attributes and relations extracted by the system**

### 7.1.16 Example Inputs and Results

This section discusses the example results of the IAA-Cyc 2 prototype that are displayed in the GUI of the above figure. The figure below shows the input document that was processed by the IAA-Cyc 2 prototype and which resulted in the output displayed in the GUI figure above.

```
MGEN XU CHENGDONG, DIRECTOR, POLITICAL DEPARTMENT (PD),
PLAAF HEADQUARTERS.  LITTLE IS KNOWN OF XU'S PAST.  HE WAS
FIRST NOTED IN PRESS REPORTS IN MARCH 1992 AS A
RESPONSIBLE OFFICER IN THE GUANGZHOU MILITARY REGION
```

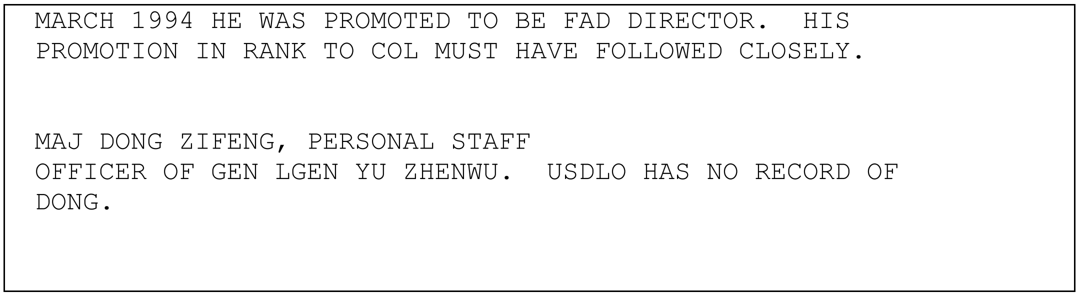
(MR). HE WAS IDENTIFIED IN JAN 1993 AS THE PD DIRECTOR OF THE GUANGZHOU AIR COMMAND, WITH THE RANK OF MGEN. HE WAS FIRST REPORTED IN MAY 1994 AS HAVING BEEN PROMOTED TO HIS PRESENT POSITION OF PD DIRECTOR, PLAAF HEADQUARTERS.

MGEN HUANG HENGMEI, DEPUTY COMMANDER CHENGDU MR AND CONCURRENT CHENGDU MR AIR FORCE COMMANDER. HUANG'S NAME FIRST APPEARED IN THE PRESS IN 1976, WHEN HE WAS DESCRIBED AS A YOUNG DEPUTY COMMANDER OF AN AIR GROUP OR SQUADRON. HE WAS NEXT SEEN IN GUANGXI IN 1984, PROBABLY ATTACHED TO THE 7TH AIR ARMY THERE. IN FEB 1991 HE SURFACED IN SHANGHAI AS COMMANDER OF THE AIR FORCE UNIT THERE, THE SHANGHAI COMMAND POST, WHICH IS A CORPS-LEVEL OR MGEN LEVEL UNIT. IN MARCH 1992 HE WAS IDENTIFIED WITH THE RANK OF MGEN. HE WAS ELECTED AS A PLA DEPUTY TO THE 8TH NATIONAL PEOPLE'S CONGRESS IN EARLY 1993. IN MID-1993 HE RELINQUISHED HIS SHANGHAI POST TO TAKE UP HIS PRESENT POSITIONS IN CHENGDU, WHICH WAS FIRST REPORTED IN JAN 1994. HIS MOST RECENT APPEARANCE IN CHENGDU WAS REPORTED IN SICHUAN RIBAO 24 FEB 1995.

MGEN ZHU YUANBIN, COMMANDER, GUANGZHOU MR'S 7TH AIR ARMY HEADQUARTERED IN NANNING, GUANGXI. A SPECIAL-GRADE AVIATOR, ZHU BECAME COMMANDER OF AN AVIATION DIVISION IN 1983. THE DIVISION WAS STATIONED ON THE YANBEI PLATEAU IN NORTHERN SHANXI PROVINCE AND SHOULD BE AN ELEMENT OF THE 10TH AIR ARMY BASED IN DATONG, SHANXI. BY FEB 1991 HE HAD BEEN TRANSFERRED TO FUJIAN PROVINCE FACING TAIWAN, TO BE A DEPUTY COMMANDER OF THE 8TH AIR ARMY. ZHU SHOWED UP AT HIS PRESENT POST IN GUANGXI IN JUNE 1994. IN AUG 1994, HE WAS CONFIRMED AS HAVING BEEN PROMOTED FROM SENIOR COLONEL TO MGEN.

COL DUAN XIAOMING, DEPUTY DIRECTOR, GENERAL OFFICE, PLAAF. DUAN IS ON USDLO RECORD'S AS BEING DIRECTOR OF THE PLAAF'S FOREIGN AFFAIRS DIVISION (FAD) UNTIL MARCH 1994. FIRST NOTED IN JAN 1980 AS AN ENGLISH INTERPRETER OF THE PLA, HE WAS IDENTIFIED AS A STAFF OFFICER OF FAD IN 1987, AND DIRECTOR OF FAD IN DEC 1991, WITH THE RANK OF COL. IT IS NOT KNOWN WHEN HE WAS PROMOTED TO HIS PRESENT POSITION. DUAN SPEAKS EXCELLENT ENGLISH AND HAS BEEN A STAFF OFFICER WORKING DIRECTLY ON ALL USAF - PLAAF CONTACTS SINCE THE MID-80'S.

COL XIN GUO, DIRECTOR, FAD, PLAAF.
THE CHINESE CHARACTERS FOR XIN'S NAME ARE NOT AVAILABLE IN USDLO FILES. HE WAS FIRST NOTED IN DEC 1992 AS A DEPUTY DIRECTOR OF FAD, WITH THE RANK OF LT COL, AND WAS REPORTED TO SPEAK FRENCH AS WELL AS ENGLISH. IN



```
MARCH 1994 HE WAS PROMOTED TO BE FAD DIRECTOR.  HIS
PROMOTION IN RANK TO COL MUST HAVE FOLLOWED CLOSELY.


MAJ DONG ZIFENG, PERSONAL STAFF
OFFICER OF GEN LGEN YU ZHENWU.  USDLO HAS NO RECORD OF
DONG.
```

**Figure 5 Input Document From Which the GUI Display Information Content was Extracted and/or Inferred**

Each row of the example Entity Attributes Table Report shows attributes of a person during a certain period of time. For example, the second row shows that Xu Chengdong was the director of the Guangzhou Air Command Political Department from some time after March 1992 until some time before May of 1994. Blank cells in a row indicate that the values for the particular column attributes are unknown. For example, the first row shows that Xu Chengdong's rank when he was an officer in the Guangzhou Military Region during March of 1992 is unknown. In these examples, the system has no prior knowledge of the persons listed in the table, so if an attribute is unknown, this means that no information on this item was extracted by the system (either because the information was not in the document(s), or the system could not extract the information or could not infer the information from what was extracted).

For the Entity Attributes Table Report GUI, the displayed time ranges pertain to the Job Position and Organization attributes, which are directly related to each other. The other attributes (e.g., Rank, Location), which are not necessarily directly related to Job Position, hold true during the time range given, without necessarily beginning or ending at the times shown. In particular, the Rank attribute is not necessarily true for the entire time range indicated. For example, Xu Chengdong may have been a Major General (MGEN) before March 1992 and after May 1994 (as is indeed shown in the third row "MGEN" value). He was however, a Major General for some time during that time range, and more specifically, a time range which includes January 1993.

Other GUI displays that are keyed on different extracted attributes (e.g., where a person was located during a period of time) are possible, though they are not currently implemented.

The values printed in red indicate that the value was implicitly expressed in the document (inferred from information extracted from the document). In the second row, the begin and end time of the position attribute is implicit since it was derived from the information that Xu Chengdong held a different position before March of 1992 (officer in Guangzhou Military Region) and a different position after May 1994 (director of the PLAAF Headquarters Political Department).

The implicit time values assume that certain positions are distinct and cannot be held concurrently. For example, if it was true that someone could hold the position of director of the Guangzhou Air Command Political Department concurrently with the position of director of the PLAAF Headquarters Political Department then the implicit End Time value shown (BEFORE May 1994) for the position of director of the Guangzhou Air Command Political Department could be incorrect. And if the position of officer in the Guangzhou Military Region is not distinct from director of the Guangzhou Air Command Political Department then the implicit End Time and Begin Time shown could be incorrect.

In the future, more knowledge will be used in determining implicit values. For example, the knowledge that "officer" is a general position value which includes the more specific "director" position and the organization "Guangzhou Military Region" includes within it as a suborganization the organization "Guangzhou Air Command Political Department" could be used to invalidate the assumption that the two positions are different (thus blocking the derivation of the implicit times related to these two positions).

There will also be in the future a confidence measure shown for both the explicit and implicit values displayed. For example, if the location of the organization Guangzhou Air Command Political Department is known to be distinct from the location of the organization PLAAF Headquarters Political Department, then that would add a measure of confidence to the assumption that the positions associated with them are both distinct and are unlikely to be held concurrently (and hence to the implicit time values shown). Also, the fact that Xu Chengdong was reported as being "promoted to" the position of director of the PLAAF Headquarters Political Department lends confidence to the two positions being distinct.

In general, knowledge about relationships between values is helpful in improving the results displayed. As discussed above, it is important to know if "officer" is more general than "director" (or is different in kind). Also, for this type of domain, it is important to know if the organization "Foreign Affair Division" is the same as the organization "PLAAF Foreign Affairs Division" (the equivalence of these two organizations is not assumed in the displayed values). Finally, it is important to know that "Squadron" and "Air Group" are not the names of any particular organizations, in contrast to other organization values, but instead express a general type of organization within the PLAAF.

In the future, the Cyc KB will be used to access knowledge for determining relationships between values and differing kinds of values so that results may be improved and the values shown may be more easily compared.

The values in each row cell of the displayed table are normalized. For example, Xu Chengdong is referred to in the document as "XU CHENGDONG", "XU", and "HE" and "HIS", all these references are normalized to "XU CHENGDONG". The references "DIRECTOR, POLITICAL DEPARTMENT (PD), PLAAF HEADQUARTERS" and "PD DIRECTOR, PLAAF HEADQUARTERS" are normalized to "PLAAF HEADQUARTERS POLITICAL DEPARTMENT DIRECTOR" (the identified position Director is then used to fill the JOB_POSITION column while the associated organization PLAAF HEADQUARTERS POLITICAL DEPARTMENT is used to fill the ORGANIZATION column). And the references "A RESPONSIBLE OFFICER IN THE GUANGZHOU MILITARY REGION" and "THE PD DIRECTOR OF THE GUANGZHOU AIR

COMMAND" are normalized to "GUANGZHOU MILITARY REGION OFFICER" and "GUANGZHOU AIR COMMAND POLITICAL DEPARTMENT DIRECTOR" respectively.

The normalization process makes use of many capabilities. As shown in the above examples, pronouns such as "he" and "his" are resolved to the persons that they refer to. Acronyms are determined from the text (e.g., POLITICAL DEPARTMENT (PD) leads to the assignment of subsequent PD's to POLITICAL DEPARTMENT). Qualification relations are recognized (e.g., as expressed in prepositional phrases such as "THE PD DIRECTOR OF THE GUANGZHOU AIR COMMAND") and form the basis of transformations to normal forms (as well as attributions).

Text expressions are also normalized that are not directly shown in the report for the purposes of system's analysis. For example, "HIS PRESENT POSITION" is normalized (i.e., the reference is resolved to through coreference processing) to "PLAAF HEADQUARTERS POLITICAL DEPARTMENT DIRECTOR". This normalization is necessary so that the sentence, "HE WAS FIRST REPORTED IN MAY 1994 AS HAVING BEEN PROMOTED TO HIS PRESENT POSITION OF PD DIRECTOR, PLAAF HEADQUARTERS" can contribute the information to the report that Xu Chengdong's current position (where "current" is shown here as the default system date 2002-02-27 when the report was run) began in May 1994; and thus his previous position ended before May 1994.

As can be seen in the report, not all of the normalizations are done correctly. For example, in the sentence, "IN MID-1993 HE RELINQUISHED HIS SHANGHAI POST TO TAKE UP HIS PRESENT POSITIONS IN CHENGDU, WHICH WAS FIRST REPORTED IN JAN 1994", the reference "HIS PRESENT POSITIONS IN CHENGDU" is not correctly normalized so that begin and end dates for positions that could be determined are not being displayed. Note also, that only one of Huang Hengmei's current positions - held concurrently - is being correctly extracted; also his position as commander of the Shanghai Command Post is being shown twice.

As regards to Huang Hengmei, knowledge was used that one may concurrently hold a PLA Deputy position (a political position) with another military position (e.g., such as Commander or Deputy Commander); thus the begin and end times of Hengmei's positions relative to his PLA Deputy position cannot be implicitly derived.

### 7.1.17 Knowledge Representation

Knowledge representation structures were designed to hold the information extracted from the messages/documents processed by the system. These structures were designed to be domain independent to the extent possible, and this effort focused on the types of information that are of highest priority to targeted end users. These information types include person, organization, and geo-political entities as well as attributes of these entities and relationships among these entities. These structures can be mapped into and represented either in a knowledge base such as the Cyc KB or a database. The structures are shown below in a table form more typical of database representation. The prototype implementation for this effort primarily used an Oracle database for speed and efficiency.

The table below presents the knowledge representation structures designed and developed as part of this IAA-Cyc 2 project. The structures will undergo refinement and enhancement to meet user and processing needs in the future.

The representational structures generally fall into four major categories:

- IBOK structures are used to hold relevant extracted information for an analyst. This information is relevant to a domain of responsibility for the analyst. This information is held in a persistent manner. It is assumed to be of value to the analyst and it is held until the analyst decides to delete or modify it.
- NLP structures are used to hold knowledge that is used during the processing of the documents by the system. This data is not regarded as having value to the analyst; it is interim data that enables the system to generate its final outputs comprised of the extracted and/or inferred information items displayed for the analyst via GUI.
- LEXICON structures are used to hold knowledge used by the system during the processing of documents. This knowledge includes vocabulary words with their attributes and processing rules for extracting attributes and relations expressed in the documents.
- SYSTEM structures are used to hold information about the documents being processed, the load groups in which the documents are included, and the status of each document as it passes through the various stages of processing.

**Table 3 Knowledge Representation Structures**

| NAME | DATA TYPE |
|---|---|
| | |
| **IBOK TABLES** | |
| | |
| IBOK_SET | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(100) |
| OWNER | VARCHAR2(50) |
| CREATE_DATE | DATE |
| MOD_DATE | DATE |

| | |
|---|---|
| COUNTRY | |
| NAME | VARCHAR2(60) |
| TYPE | VARCHAR2(30) |
| CAPITAL | VARCHAR2(60) |
| TRANS_REGION_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| TRANSNATIONAL_REGION | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(100) |
| TYPE | VARCHAR2(30) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| ORGANIZATION | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(90) |
| TYPE | VARCHAR2(30) |
| SUBTYPE | VARCHAR2(30) |
| ECHELON | VARCHAR2(90) |
| SERVICE_BRANCH | VARCHAR2(90) |
| SPECIALTY | VARCHAR2(90) |
| PARENT_ORG_ID | NUMBER(9) |
| ASSOC_FACILITY_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| FACILITY | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(90) |
| TYPE | VARCHAR2(30) |
| ASSOC_ORG_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| PERSON | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(90) |
| TYPE | VARCHAR2(30) |
| FIRST_NAME | VARCHAR2(40) |
| SECOND_NAME | VARCHAR2(40) |
| LAST_NAME | VARCHAR2(40) |
| FAMILY_NAME | VARCHAR2(60) |

| | |
|---|---|
| GENDER | VARCHAR2(1) |
| BIRTH_DATE_ID | NUMBER(9) |
| DEATH_DATE_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| ARTIFACT | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(90) |
| TYPE | VARCHAR2(30) |
| SUPER_CLASS | VARCHAR2(30) |
| SUB_CLASS | VARCHAR2(30) |
| HOLDER_ID | NUMBER(9) |
| HOLDER_TYPE | VARCHAR2(30) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| TIME_VAL | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| NORMALIZED_TIME | DATE |
| MIL | NUMBER(4) |
| CEN | NUMBER(4) |
| DC | NUMBER(4) |
| YR | NUMBER(4) |
| MON | NUMBER(4) |
| DY | NUMBER(4) |
| HR | NUMBER(4) |
| MT | NUMBER(4) |
| SEC | NUMBER(4) |
| DESCRIPTION | VARCHAR2(90) |
| | |
| TIME_RANGE | |
| OBJECT_ID | NUMBER(9) |
| TIME_TYPE | VARCHAR2(30) |
| VAL_TYPE | VARCHAR2(30) |
| SECOND_TIME_ID | NUMBER(9) |
| REL_INFO_TO_FIRST_ID | NUMBER(9) |
| REL_INFO_TO_FIRST_TYPE | VARCHAR2(30) |
| REL_INFO_TO_SECOND_ID | NUMBER(9) |
| REL_INFO_TO_SECOND_TYPE | VARCHAR2(30) |
| DESCRIPTION | VARCHAR2(90) |
| DURING_TIME_ID | NUMBER(9) |
| | |

| | |
|---|---|
| LOCATION | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(100) |
| TYPE | VARCHAR2(30) |
| SUBTYPE | VARCHAR2(30) |
| CONFIDENCE | NUMBER(9) |
| PRECISION_VAL | NUMBER(9) |
| COUNTRY | VARCHAR2(60) |
| CITY | VARCHAR2(60) |
| POLITICAL_DIVISION_1 | VARCHAR2(50) |
| POLITICAL_DIVISION_TYPE_1 | VARCHAR2(50) |
| POLITICAL_DIVISION_2 | VARCHAR2(50) |
| POLITICAL_DIVISION_TYPE_2 | VARCHAR2(50) |
| POLITICAL_DIVISION_3 | VARCHAR2(50) |
| POLITICAL_DIVISION_TYPE_3 | VARCHAR2(50) |
| LATITUDE_1 | VARCHAR2(7) |
| LATITUDE_2 | VARCHAR2(7) |
| LONGITUDE_1 | VARCHAR2(8) |
| LONGITUDE_2 | VARCHAR2(8) |
| LATITUDE_DEGREES | NUMBER(9) |
| LATITUDE_DIRECTION | VARCHAR2(1) |
| LONGITUDE_DEGREES | NUMBER(9) |
| LONGITUDE_DIRECTION | VARCHAR2(1) |
| RELATIVE_ID | NUMBER(9) |
| | |
| DISTANCE_RANGE | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(20) |
| CONFIDENCE | NUMBER(9) |
| MINIMUM_DISTANCE | NUMBER(10) |
| MAXIMUM_DISTANCE | NUMBER(10) |
| START_DIRECTION | NUMBER(10) |
| STOP_DIRECTION | NUMBER(10) |
| | |
| META_INFORMATION | |
| OBJECT_ID | NUMBER(9) |
| EXTERNAL_SOURCE | VARCHAR2(250) |
| DOCUMENT_ID | NUMBER(9) |
| JUDGEMENT | VARCHAR2(60) |
| JUDGEMENT_BASIS_TYPE | VARCHAR2(60) |
| JUDGEMENT_BASIS_1 | VARCHAR2(60) |
| JUDGEMENT_BASIS_2 | VARCHAR2(60) |
| CONFIDENCE_SOURCE | VARCHAR2(60) |
| CONFIDENCE | NUMBER(9) |

| | |
|---|---|
| REPORT_DATE_ID | NUMBER(9) |
| REPORTED_BY_ORG | NUMBER(9) |
| REPORTED_BY_PERSON | NUMBER(9) |
| LOCATION_ID | NUMBER(9) |
| LOCATION_META_TYPE | VARCHAR2(30) |
| TIME_HOLDS_ID | NUMBER(9) |
| TIME_HOLDS_META_TYPE | VARCHAR2(30) |
| | |
| ATTRIBUTION_INSTANCE | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| SEMANTIC_ID | NUMBER(9) |
| SEMANTIC_TYPE | VARCHAR2(30) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| ATTRIBUTION | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| INSTANCE_ID | NUMBER(9) |
| ATTRIBUTE_ID | NUMBER(9) |
| HOLDER_ID | NUMBER(9) |
| HOLDER_TYPE | VARCHAR2(30) |
| META_INFO_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| ATTRIBUTE | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(90) |
| TYPE | VARCHAR2(30) |
| PREVIOUS_ID | NUMBER(9) |
| NEXT_ID | NUMBER(9) |
| IBOK_SET_ID | NUMBER(9) |
| | |
| RELATIONSHIP | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| TEXT | VARCHAR2(1000) |
| INITIATOR_EXPR_ID | NUMBER(9) |
| RECEIVER_EXPR_ID | NUMBER(9) |
| DIRECTION | VARCHAR2(10) |
| NATURE | VARCHAR2(60) |
| OUTPUT_CODE | VARCHAR2(30) |
| REL_LEXICON_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |

| | |
|---|---|
| **NLP TABLES** | |
| | |
| TEXT_REFERENCE | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| TEXT | VARCHAR2(1000) |
| SYNTACTIC_TYPE | VARCHAR2(30) |
| SEMANTIC_TYPE | VARCHAR2(30) |
| SEMANTIC_ID | NUMBER(9) |
| HEAD_LEXICON_MATCH | VARCHAR2(250) |
| ROOT_FORM | VARCHAR2(250) |
| NORMAL_FORM | VARCHAR2(250) |
| GENDER | VARCHAR2(1) |
| NUMERIC_FEATURE | VARCHAR2(30) |
| ANIMACY_FEATURE | VARCHAR2(30) |
| GRAMMATICAL_FUNCTION | VARCHAR2(30) |
| COREF_EQUIV_CLASS_ID | NUMBER(9) |
| COREF_SUPER_CLASS_ID | NUMBER(9) |
| COREF_SUB_CLASS_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| CROSS_EQUIV_CLASS_ID | NUMBER(9) |
| | |
| COREFERENCE_CLASS | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| DOCUMENT_ID | NUMBER(9) |
| ORDINAL_NUM | NUMBER(9) |
| NORMAL_FORM | VARCHAR2(250) |
| EXEMPLAR_ID | NUMBER(9) |
| EXEMPLAR_TYPE | VARCHAR2(20) |
| COREF_SUPER_CLASS_ID | NUMBER(9) |
| COREF_SUB_CLASS_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| SEMANTIC_TYPE | VARCHAR2(30) |
| SEMANTIC_ID | NUMBER(9) |
| | |
| EVENT | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(60) |
| TYPE | VARCHAR2(30) |
| ACTOR_ID | NUMBER(9) |
| ACTOR_TYPE | VARCHAR2(30) |
| AFFECTED_ID_1 | NUMBER(9) |
| AFFECTED_TYPE_1 | VARCHAR2(30) |

| | |
|---|---|
| AFFECTED_ID_2 | NUMBER(9) |
| AFFECTED_TYPE_2 | VARCHAR2(30) |
| VERB_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| | |
| VERB | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| TEXT | VARCHAR2(100) |
| NORMAL_FORM | VARCHAR2(100) |
| ROOT_FORM | VARCHAR2(100) |
| TENSE | VARCHAR2(30) |
| ASPECT | VARCHAR2(30) |
| | |
| STATE | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(60) |
| TYPE | VARCHAR2(30) |
| SUBJECT_ID | NUMBER(9) |
| SUBJECT_TYPE | VARCHAR2(30) |
| STATE_CONDITION | VARCHAR2(60) |
| STATE_CONDITION_TYPE | VARCHAR2(30) |
| VERB_ID | NUMBER(9) |
| META_INFO_ID | NUMBER(9) |
| | |
| EXPRESSION_LINK | |
| OBJECT_ID | NUMBER(9) |
| INFORMATION_ID | NUMBER(9) |
| INFORMATION_TYPE | VARCHAR2(30) |
| INFORMATION_NAME | VARCHAR2(30) |
| EXPRESSION_ID | NUMBER(9) |
| EXPRESSION_TYPE | VARCHAR2(30) |
| DOCUMENT_ID | NUMBER(9) |
| BEGIN_TEXT_OFFSET | NUMBER(9) |
| END_TEXT_OFFSET | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| | |
| **LEXICON TABLES** | |
| | |
| LEXICON_SET | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(250) |
| OWNER | VARCHAR2(50) |
| CREATE_DATE | DATE |
| MOD_DATE | DATE |

| | |
|---|---|
| TERM_LEXICON | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| TERM | VARCHAR2(100) |
| ROOT_FORM | VARCHAR2(100) |
| SEMANTIC_ID | NUMBER(9) |
| SEMANTIC_TYPE | VARCHAR2(30) |
| GENDER | VARCHAR2(1) |
| NUMERIC_FEATURE | VARCHAR2(20) |
| ANIMACY_FEATURE | VARCHAR2(30) |
| ARGUMENT | VARCHAR2(50) |
| ARGUMENT_TYPE | VARCHAR2(30) |
| ARGUMENT_MARKER | VARCHAR2(50) |
| LEXICON_SET_ID | NUMBER(9) |
| KB_ID | VARCHAR2(50) |
| | |
| ATTRIBUTION_LEXICON | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(20) |
| RANGE | VARCHAR2(20) |
| PATTERN_STRING | VARCHAR2(250) |
| ROOT_FORM | VARCHAR2(250) |
| ATTRIBUTION_ID | NUMBER(9) |
| OUTPUT_TABLE_1 | VARCHAR2(30) |
| OUTPUT_FIELD_1 | VARCHAR2(30) |
| OUTPUT_TABLE_2 | VARCHAR2(30) |
| OUTPUT_FIELD_2 | VARCHAR2(30) |
| LEXICON_SET_ID | NUMBER(9) |
| | |
| RELATIONSHIP_LEXICON | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| PATTERN_STRING_1 | VARCHAR2(250) |
| PATTERN_STRING_2 | VARCHAR2(250) |
| PATTERN_STRING_3 | VARCHAR2(250) |
| ROOT_FORM | VARCHAR2(250) |
| DIRECTION | VARCHAR2(10) |
| NATURE | VARCHAR2(60) |
| OUTPUT_CODE | VARCHAR2(30) |
| LEXICON_SET_ID | NUMBER(9) |
| | |
| FEATURES | |
| OBJECT_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |

| | |
|---|---|
| TABLE_OBJECT_ID | NUMBER(9) |
| TABLE_NAME | VARCHAR2(30) |
| FEATURE_NAME | VARCHAR2(30) |
| FEATURE_VALUE | VARCHAR2(90) |
| | |
| **SYSTEM TABLES** | |
| | |
| LOAD_QUEUES | |
| OBJECT_ID | NUMBER(9) |
| FILE_IN | VARCHAR2(250) |
| SOURCE | VARCHAR2(250) |
| STATUS | VARCHAR2(30) |
| PROCESS_HOST | VARCHAR2(40) |
| ENQUEUE_DATE | DATE |
| MOD_DATE | DATE |
| LOAD_GROUP_ID | NUMBER(9) |
| TYPE | VARCHAR2(30) |
| PROCESS_AFTER_QUEUE_ID | NUMBER(9) |
| | |
| LOAD_GROUP | |
| OBJECT_ID | NUMBER(9) |
| NAME | VARCHAR2(250) |
| OWNER | VARCHAR2(50) |
| CREATE_DATE | DATE |
| MOD_DATE | DATE |
| TYPE | VARCHAR2(30) |
| WDN_PROCESS | NUMBER(1) |
| CDN_PROCESS | NUMBER(1) |
| REL_PROCESS | NUMBER(1) |
| ATTR_PROCESS | NUMBER(1) |
| | |
| DOCUMENT | |
| OBJECT_ID | UMBER(9) |
| TITLE | VARCHAR2(100) |
| TYPE | VARCHAR2(30) |
| LOAD_GROUP_ID | NUMBER(9) |
| FILENAME | VARCHAR2(250) |
| SOURCE | VARCHAR2(250) |
| AUTHOR | VARCHAR2(100) |
| DOC_DATE_ID | NUMBER(9) |
| DOC_DATE_META_TYPE | VARCHAR2(20) |
| PROCESS_DATE | DATE |
| CONTEXT_DATE_ID | NUMBER(9) |
| CONTEXT_DATE_META_TYPE | VARCHAR2(20) |
| CLASSIFICATION | VARCHAR2(30) |

| | |
|---|---|
| LOCATION_ID | NUMBER(9) |
| LOCATION_META_TYPE | VARCHAR2(20) |
| WDN_PROCESS_DATE | DATE |
| CDN_PROCESS_DATE | DATE |
| REL_PROCESS_DATE | DATE |
| ATTR_PROCESS_DATE | DATE |
| | |

### 7.1.17.1 Database Design Notes

#### *7.1.17.1.1 Temporal Information*

The temporal structures were designed to accommodate the wide variety of natural language expressions for dates and times, including time values and ranges. The intent is to accommodate the semantics of a great variety of natural language temporal expressions, beyond dates/times that can be normalized to a specific instance in time. Examples include the following:

- “prior to 1996”,
- “sometime in March of 1999”, and
- “between 1996 and 1999.”

The TIME_VAL structure was designed to represent an instance in time whereas the TIME_RANGE structure was designed to represent absolute or relative times of various types (ranges, before a specified time, after a specified time, descriptive expressions for dates/times, etc.). The fields of “TIME_RANGE” point to one or more records of the “TIME_VAL” table that hold partial or complete information about single time values. The two tables together accurately and flexibly represent complete or incomplete temporal information. Extracted or information-level body-of-knowledge (IBOK) records include pointers to stored time ranges to represent metadata concerning the time period during which an information item holds true.

#### *7.1.17.1.2 Spatial Information*

The LOCATION structure is designed to represent named locatives of various types including geopolitical entities such as cities, counties, villages, countries, provinces, among others. The LOCATION structure is also designed to represent the location of objects that are absolute, such as at a specified lat-long, or relative to the location of another object with a direction such as north or south. The DISTANCE_RANGE is designed to represent distance ranges expressed by a phrase such as "20 to 30 miles north of Toledo".

#### 7.1.17.2 Database Scripts for System Administrator Support

Scripts were developed to create, maintain, delete, reset, and maintain the IAA-Cyc database including users, table space, database tables, sequences, and indexes. The following table summarizes the available scripts.

**Table 4 Database Scripts**

| # | DESCRIPTION |
|---|---|
| 1 | Creates the owner of the IAA-Cyc database instance. |
| 2 | Adds a user to the IAA-Cyc database instance. |
| 3 | Creates the indexes in the IAA-Cyc database. |
| 4 | Creates the NULL table rows in the IAA-Cyc database. NULL table rows are used to avoid outer joins in queries. |
| 5 | Creates the sequences for in the IAA-Cyc database. Sequences are used to generate unique OBJECT_IDs for tables. |
| 6 | Creates synonyms for the IAA-Cyc database. Synonyms give a short name to a table for the user's convenience. |
| 7 | Creates a file of size parameters included by other scripts listed here. |
| 8 | Saves a description of each of the IAA-Cyc tables. |
| 9 | Drops the IAA-Cyc  indexes. |
| 10 | Drops the IAA-Cyc  sequences. |
| 11 | Drops the IAA-Cyc  tables. |
| 12 | Grants permissions to an IAA-Cyc user for the IAA-Cyc tables. |

### 7.1.18 Conformity with Expectations and Confidence Scoring

As part of this project, we researched the manner in which the Cyc KB could be applied to checking the consistency and expectedness of extracted information with respect to other known information.

We investigated ways to define the knowledge and methods to be represented in the Cyc KB and used in the IAA-Cyc 2 software and capabilities demonstrations. The majority of this research was performed by Cycorp and is discussed in Section 7.2.  The following topics were researched by Veridian in collaboration with Cycorp:

- Representation of expectedness, degrees of expectedness, and rules to reason about the conformity of extracted information to expectations.
- The definition and representation of expectations in areas of particular interest. An example is the direction of career paths, especially in the Chinese military.
- Representation of temporal granularity and rules to determine and reason about the granularity of temporal meta-information associated with extracted information.

The manner in which the KB should encode knowledge concerning these topics was investigated. The use of knowledge concerning these topics to detect "real-world" anomalies and errors in the extraction processes was also investigated.

This investigation included the inference of person attributes from explicitly extraction information. This work focused on the following three areas of knowledge:

1. Given a job position in Chinese Military (e.g., Commander), what is the expected rank (e.g., Major General)?
2. Given two job positions, are they expected to be held concurrently or not (e.g., Commander and PLA Deputy)?
3. Given two job positions, what are the expected intervening positions(s), based upon a hierarchy of positions?

We investigated and developed a preliminary design for the generation of confidence scores for information items based, in part, on the degree to which the information conforms with expectations. In the area of attribute association, our preliminary design accommodates the use of the following sources of evidence in judging the confidence of an attribute associated with a person (e.g., a person has a certain position at a certain time) or organization:

- Degree of confidence in the extraction engine or software component.
- Reliability of the source/author of the document.
- The nature of the natural language used to express the attribute, such as being stated as a definite assertion (e.g., "Deng Li is a favorite of …") versus as a possibility (e.g., "Deng Li is believed to be a favorite of …").
- Comparison with other extracted results from the same document or other documents.
- Compatibility with expectations encoded in the knowledge base about general or typical event sequences or scripts.

The confidence score of an extraction result will be based upon the syntactic context of the attribute and the person or organization associated with it. Heuristic estimates will be used for confidence based upon an evaluation of results. For example, position attributes that are within the same noun phrase as a person (with no other positions or persons) will be given a higher confidence scores than those that span a clause. And within a noun phrase, those position attributes that immediately precede or follow in appositive constructions will be given higher confidence scores.

Extraction results will be compared within a document in order to determine support for a particular result. For example, if a particular information item is extracted from multiple locations within a document, then this will increase the confidence score for the item. The more certain results will also be “propagated” within a document. Results within a document may be related in other ways that add or subtract support. For example, there may be two different results that specify a person’s position at a certain time. If these two different positions are unlikely to be held at the same time (per information from the KB), then the position that is more likely, based upon extraction confidence score, will be favored.

The database of extracted results will be checked for other facts that concern the person or organization to which an extracted attribute could pertain. This “instance-level” knowledge will provide support for an extracted result. For example, if the same result was extracted from another document, then the confidence score of the extracted result would be raised if the second document was from a different source, or if the document was from the same source but the extraction result has a higher extraction confidence score.

The knowledge base will be used for “type-level” compatibility checking of extraction results. For example, if a person was associated with a position, the KB would be checked to determine whether the type of position conforms to the type of person. For example, if, by the extraction of a rank title for a person, the person was classified as a MilitaryPerson, and the position was classified as a MilitaryPosition such as “deputy commander”, then the confidence score would be increased due to the rule that a MilitaryPerson is expected to hold a MilitaryPosition.

The following categories of compatibility knowledge were incorporated into the KB design:

- Type Compatibility such as:
    - compatibility of type of person/organization with type of the attribute.
    - compatibility of a type of position (e.g., professional) with a type of education (e.g., professional degree).
    - compatibility of a type of position (e.g., military position) with a type of employer (e.g., military organization).
- Location Compatibility such as:
    - compatibility of a person's address with their employer's address.
- Time Compatibility such as:
    - compatibility of two different attribute values or types holding at the same time.
    - compatibility of a current position with a previous position as determined from expectations based upon a KB hierarchy of positions.

The development of this knowledge in the KB will be based upon both:

- Common sense principles (e.g., a person is unlikely to live in a country that is different from the one in which they work, making allowances for permanent versus temporary residence and military postings) and
- Domain specific information (much of which was developed as part of the IAA-Cyc 1 project).

This area of expectations representation and measuring the conformity of extracted information with expectations is an on-going area of work. The Cyc KB ontological engineering accomplishments are discussed in detail in Section 7.2.

### 7.1.19 Testing

Formal test procedures have been developed as a part of this phase of the IAA-Cyc project. They will be finalized and used on the next phase of the IAA-Cyc project. The test procedures are based, in part, on guidelines and standards developed by the Message Understanding Conference (MUC) program and the Automated Content Extraction (ACE) program, with modifications and extensions to the standards to accommodate the IAA-Cyc 2 tag set.

An initial test suite of documents has been selected from Government sources. The test suite will be augmented to make it as representative of the target documents as possible to provide for the best estimates of how the system will perform under operational conditions.

The initial test suite has been annotated using the defined standard to create a set of "ground truth" answer-keys. The process of annotating the keys revealed areas in which the IAA-Cyc standards/guidelines needed refinement, and, as a result, the process resulted in some iterative refinement of these standards/guidelines.

The test documents will be processed using the IAA-Cyc system and the results will be compared to the "ground truth" keys and scored.

This testing task has leveraged the efforts and results of other projects being performed by Veridian.

### 7.1.20 Final Technical Interchange Meeting and Demonstration

The final Technical Interchange Meeting (TIM) for IAA-Cyc 2 was held at AFRL Rome Research Site on 21 March 2002. Project accomplishments were reviewed at the TIM, as well as the goals and objectives for phase 3. The final IAA-Cyc 2 software release was demonstrated and feedback was gathered from the TIM participants.

## 7.2 Cyc KB Ontological Engineering

### 7.2.1 Objectives

The goal of the Cyc KB ontological engineering work in the IAA-Cyc 2 project was to demonstrate the utility of the Cyc ontology and inferencing rules, not just for information extraction but for reasoning about and specifically for anomaly checking against extracted information that had been stored in a structured database format. The motivation for doing this was twofold, stemming on both hands from the significant information overload faced by contemporary analysts. On the one hand, a strong premium is believed to attach to being able to detect and flag patterns in real-world data reflecting genuinely anomalous conditions of interest to an analyst, e.g., an individual who "rises through the ranks" of an organization much more quickly than would ordinarily be expected. At the same time, a high degree of utility would also attach to a system that could analyze a corpus of extracted data for anomalies that must reflect flaws in either data collection methods or in the extraction process itself. Speaking generally, the same desideratum underlies both objectives: namely, the ability to compare the information encoded in a structured database to general expectations in such a way as to detect divergence between expectations and represented facts. This ability is what we have endeavored to implement, in an initial, scalable form, in Cyc in the IAA-Cyc 2 project.

### 7.2.2 General Approach

The main challenge to using standard CycL inference rules for purposes of anomaly detection is quite fundamental and has to do with the implementation of inference in a rule-based system. Suppose it is the case that "other things being equal" or "under normal circumstances", something that satisfies conditions C1...Cn will satisfy E (where C1...CN and E are open formulae featuring a variable into which an individual may be instantiated, but there are a few abnormal things which satisfy C1...Cn but which don't satisfy E, and it's precisely these abnormal cases that we want to detect. If one is attempting to do conformity-checking, it will not do simply to encode this expectation with a rule (-> (C1&...&Cn) E) with universal quantification over the variable position, even if -> is interpreted as a defeasible or non-monotonic operator and not the material conditional. This is because, if an individual satisfies (C1&...&Cn), then, absent any positive evidence to the effect that the individual does *not* satisfy E, this will be concluded, effectively preventing the realization in the knowledge base of the very state of affairs one would like to detect, namely, a situation in which the data contains the information that an individual satisfies C1...Cn, but not the expected information that the individual also satisfies E.

The solution is to introduce an explicit expectation operator taking a propositional argument, and formulate the conformity-checking rule with this operator wrapping the consequent, thus: (-> (C1&...&Cn) (expec E)). It then becomes possible to check whether there are any individuals which are expected to satisfy E but which cannot be proved to satisfy it, and more generally, whether there are any expectations in the reasoning domain which are unsatisfied.

This general approach has been taken w.r.t. all of the anomaly-checking work undertaken for IAA-Cyc up to the present.

### 7.2.3 Focus

The field of conformity/anomaly checking affords a very broad scope field of opportunity for work. Simply specifying anomalies that are of interest to the analyst to the degree necessary to admit encoding in expectation-checking rules can be a non-trivial task, and there are many different avenues to be explored. For the initial work for IAA-Cyc 2, we elected to concentrate on occupational positions. Critically, we elected *not* to directly address the question of whether the type of conformity checking involved was intended to focus on: detection of real-world abnormalities vs. detection of faulty data. In all of the cases specified below, it could be either, and indeed, it would probably take the intervention of a human analyst to make the determination.

1. given a job position in the Chinese Military (e.g. Commander) encode the expected rank (e.g. Major General) associated with that position and use this for detecting cases where an individual in a given position has a higher-than-expected rank.

2. given 2 job positions, encode whether they are expected to be held concurrently or not, and detect cases when two positions are being held concurrently that are not expected to be so.

3. given 2 job positions on a "career path" in the Chinese armed forces, encode what are the expected intervening positions on that path and detect cases where an individual is not known to serve in all of these positions; i.e., where an individual seems to be diverging from an expected career path.

As comparatively simple as these tasks seem, and although versions of all of these inference types have in fact been successfully implemented using hand-entered test data in the Cyc knowledge base, they present nontrivial and extremely instructive challenges for a rule-based reasoning system like Cyc. A brief overview of some of the principle technical issues that had to be dealt with in getting the inferences to work serves to illustrate both the promise of conformity checking in a rule-based system and the main difficulties that must be overcome in order to implement it.

### 7.2.4 Issues Encountered

#### 7.2.4.1 Conceptual Issues

##### *7.2.4.1.1 What to detect: errors vs. real world anomalies*

As noted, a central background issue attaching to all of the IAA ontology work has to do with whether the anomaly detecting work is aiming to trap for unusual situations in the real world that are reflected in patterns in the extracted data, or is trying to find patterns in the extracted data that reflect either faulty collection mechanisms or faulty extraction mechanisms.

Data that violates extremely strong expectations probably is indicative of faulty extraction or data collection. Violations of "strong expectations" can of course take the form of data from which we can infer outright logical contradictions, but it might also take the form of information that is in radical violation of common sense knowledge, such as the datum that Osama Bin Laden has the occupational position "Pope". In a system like Cyc, such common sense knowledge might take the form of encoded argument constraints or disjointness so that contravening information would violate well-formedness constraints. Violations of weaker expectations, such as we might try to encode with expectation-checking rules, are ambiguous with respect to interpretation: they could be indicative of faulty data, or they could reflect an interesting abnormality in the real world. The weaker the expectations in question are, the greater this ambiguity becomes.

One feature that this obviously suggests a need for is an ability to represent gradation in the strength of expectations.

### 7.2.4.2 Integration Issues

#### *7.2.4.2.1 DB and KB*

Historically, inference in Cyc has employed conditional rules of the form:

```
 (implies
   (ANTECEDENT CLAUSES)
   (CONSEQUENT CLAUSE))
```

The actual implementation uses resolution proof with heuristic search of the resultant tree of literals: this is less important than the fact that there is a fundamental pressuposition to the effect that the ground atomic formulas that bind to the antecedent of the rule in question (or rules, in the case of multiple backchain inference) are actually in the Cyc Knowledge Base.

This obviously poses a potential problem for cases where Cyc needs to reason about the contents of very large databases, such as might be produced by an extraction engine working on any sizeable body of text, to the extent that it may not be practical to translate all of the database tuples into CycL assertions and import them into Cyc. An ongoing effort is currently underway at Cycorp to deal problems of this type, to the extent that Cyc would be able to access and reason about information in a distinct distributed back-end data store.

This work is still in its initial phase, though, and we have not availed ourselves of any results of it in this phase of the project.

### 7.2.4.3 Inference Issues

The three main inference issues encountered in the year two work have to do, respectively, with reasoning from within intensional context, temporal qualification, and truth maintenance.

#### *7.2.4.3.1 Intensionality*

The heavy reliance on the "modal" expectation predicates **expected-ToBe** and **expectationsConcerning** (they are "modal" in the strict sense of taking propositional arguments) exposes both a significant OE issue and a fundamental limitation of present Cyc inference. On the inference side, although it is desirable that Cyc be able to prove, given that **(expected-ToBe PROPOSITION)** and **(implies PROPOSITION IMPLIED-CONCLUSION)**, that **(expected-ToBe IMPLIED-CONCLUSION)**, it is in practice often quite difficult to achieve this result. Indeed, it is logically impossible that we should ever be in a position of being able to guarantee that we can achieve it for arbitrary cases: the fact that CycL is a semi-decideable language means that there is no decision procedure that can tell us that a putative conclusion can*not* be inferred from a given set of premises. Failure on the part of the inference engine to infer something means just that: that the inference engine failed to find a proof, using its extant HL modules and heuristic search procedure. It doesn't mean, and can't, mean that no proof exists.

That said, Cyc's general ability to propagate the modal wrapper of a propositional argument to the logical consequences of that argument is still quite limited, apart from a few particular cases with HL reasoning support. In general it is not safe, for present inference purposes, ever to count on this ability, even where the inference looks reasonably straightforward to a human. Thus, to take a simple example, the fact that Cyc can prove, say,

```
(expected-ToBe
  (thereExistsAtLeast 50 ?X
     (rank-Military ?X MajorGeneral-Rank)))
```

does not guarantee that Cyc can prove

```
(expected-ToBe
  (thereExists ?X
     (rank-Military ?X MajorGeneral-Rank)))
```

even though this seems obvious. There's also a corollary to this point that is even more significant for current anomaly checking. Even given that Cyc knows **(expected-ToBe PROPOSITION)**, it is often hard for Cyc to tell when PROPOSITION is entailed by extant propositions in the knowledge base in cases where PROPOSITION itself has not been explicitly asserted. This again is a consequence of fundamental limitations on our ability to check implications from within intensional contexts, in this case from within the context of a CycL query. To see the consequences that this has, suppose that Cyc can infer the expectation

```
(expected-ToBe
  (thereExists ?X
     (rank-Military ?X MajorGeneral-Rank)))
```

and suppose further that what is known explicitly in the reasoning domain is that

```
(rank-Military ZhuYuanbin MajorGeneral-Rank)
```

Now, in fact, the inferred expectation is satisfied: there *is* someone in the knowledge base who has the rank **MajorGeneral-Rank**. However, if we query to find all of the unsatisfied expectations in the reasoning context via a query of the form:

```
(and
  (expected-ToBe ?WHAT)
(unknownSentence ?WHAT))
```

we will get back

```
(thereExists ?X
   (rank-Military ?X MajorGeneral-Rank))
```

as an answer because the system will not be able to check that (rank-Military ZhuYuanbin MajorGeneral-Rank) entails this within the context of the query.

#### *7.2.4.3.2 Temporal qualification issues*

The second of the targeted conformity checks described above implicates a strong temporal reasoning component insofar as 'concurrently' accommodates a certain vagueness of definition. We cannot reasonably expect that an extraction will produce the information that an agent holds two particular positions simultaneously, and indeed, even if this were the case, just checking for this condition alone would not be adequate. What really requires checking is whether or not an agent is recorded as holding two positions within a time frame that is sufficiently small as to be "suspicious".

What is meant by "sufficiently small" of course varies with respect to circumstance, particularly the general type of position (are we talking roles in a military organization, a commercial organization, or what?). For the purposes of the exercise, we have seen fit to define a 'nonconcurrent positions' predicate with an argument that specifies the granularity of the time interval. Two versions of this were created, one that uses **OrganizationalPosition**s and one that uses **OccupationType**s: both help to illustrate some of the problems temporal qualification can pose for conformity checking, and what can be done about them. The expression (nonConcurrentPositions-OrganizationalPosition ORG POS1 POS2 DURATION) means that we would not expect to find the same individual holding POS1 and POS2 in ORG within the space of a time-frame demarcated by DURATION. More specifically, (nonConcurrentPositions-OrganizationalPosition ORG POS1 POS2 DURATION) means that if it is the case that (holdsIn TEMP1 (POS1 ORG AGENT)) and (holdsIn TEMP2 (POS2 ORG AGENT)), the expectation is that the duration of (TimeIntervalBetweenFn TEMP1 TEMP2) is greaterThan DURATION, with the definitional rule formulated thusly:

```
(implies
 (and
  (duration (TimeIntervalBetweenFn ?TEMP1 ?TEMP2) ?DUR)
  (holdsIn ?TEMP1 (?POS1 ?ORG ?PERSON))
  (holdsIn ?TEMP2 (?POS2 ?ORG ?PERSON))
  (nonConcurrentPositions-OrganizationalPosition ?ORG ?POS1 ?POS2 ?TIME-FRAME))
 (expectationConcerning ?PERSON (greaterThan ?DUR ?TIME-FRAME)))
```

The relation nonConcurrentPositions-OccupationType is defined analogously, but using **OccupationType** in place of **OrganizationalPosition**.

Such a relation can be used, in principle, to check whether an agent is represented in a data set as holding two positions within the scope of a specified time frame: if it is known that

**(nonConcurrentPositions-OrganizationalPosition ORG POS1 POS2 DURATION)** and an agent, AGENT, say,is represented as holding **POS1** in **ORG** in time frame **TEMP1** and **POS2** in **ORG** in **TEMP2**, then, where **INTERVAL** is the time interval between **TEMP1** and **TEMP2** the query

```
(and
 (expectationConcerning AGENT ?WHAT)
 (unknownSentence ?WHAT))
```

will return **(greaterThan DURATION INTERVAL)** in the case where **INTERVAL** is in fact equal to or greater than the specified duration.

The only problem in this case is calculating the time interval between two arbitrarily specified time intervals. This can be quite a challenge, given that interval-based reasoning in Cyc is still quite limited. Up to the present, it has been necessary to rely upon comparatively ad hoc rules specific to certain date types, e.g., dates expressed using the individual denoting function **YearFn**:

```
(duration
  (TimeIntervalBetweenFn (YearFn ?START-YEAR) (YearFn ?END-YEAR))
  (YearsDuration
    (DifferenceFn (DifferenceFn ?END-YEAR ?START-YEAR) 1)))
```

These significantly limit the current utility of this and related approaches to checking conformity with temporal expectations. A truly comprehensive schema for reasoning about the durations of intervals between arbitrary dates is probably best implemented in Cyc with partial code support. Such a schema may in fact soon be implemented as an adjunct to an overhaul of the Cyc temporal reasoning system that involves replacing the older holdsIn convention with a more systematically axiomatized modal tense logic. Such a system will ultimately require revision of all reasoning rules involving **holdsIn**, including the definitions of the nonconcurrent position predicates.

#### *7.2.4.3.3 Truth Maintenance Issues*

The present round of OE work in year two has not focused directly on contradiction checking, even though, in addition to being able to detect cases where expectations go unsatisfied by extracted data, it would be very useful to be able to identify cases where the data directly contradicts expectations. Unfortunately, our ability to do this in reasoning at the present time is fundamentally restricted by the way that truth maintenance in the Cyc KB is handled. It is certainly true that the Cyc KB is outfitted with extremely effective wff-checking diagnostics that can readily detect well-formedness violations, and Cyc is also extremely good at detecting and signaling error conditions on violations of semantic constraints: for example, argument constraints on predicates, or disjointess constraints on collections. Probably, Cyc is one of the best systems available today at rejecting inputs that conflict with previously asserted knowledge. However, and somewhat paradoxically, this has the effect of making use of the inference engine for certain kinds of reasoning about contradictions *within* the knowledge base extremely hard.

Strictly speaking, if a proposition that someone is trying to assert contradicts a fact that is already asserted in the knowledge base and visible in the reasoning domain where the user is trying to assert the new fact, one of two things will happen. If the proposition contradicts a fact that is monotonically true or the result of a rule that is monotonically true, the new fact will simply be rejected out-of-hand with an error message of the appropriate type: it can never be asserted to the knowledge base. Conversely, if the prior fact is defeasible or the conclusion of a rule that is defeasible, it will simply be retracted in favor of the new one. Where defeasible rules are concerned, this truth maintenance scheme is what makes it possible for us to develop 'pro' and 'con' arguments for a given proposition within a single reasoning domain, and it also supports and is supported by the Cyc microtheoretic hierarchy system, whereby information that would otherwise conflict can be sequestered into epistemically disjoint reasoning domains. However, it also makes it next to impossible to do anything like proof by reduction to absurdity within a single reasoning domain, since the truth maintenance system will assiduously fight any attempt to simultaneously assert contradictory propositions within the same context. This means in turn that any kind of contradiction checking in Cyc, within the context of expectation reasoning or anything else, is essentially limited to a form of error handling that utilizes well-formedness checking diagnostics and semantic constraint checking. That is, there can at this time be no question of a utility that first loads database content into Cyc in the form of assertions and *then* attempts to check for contradictions between existing knowledge in the reasoning domain and what has been loaded. Rather, we must think in terms of utilities which attempt to analyze the error detections that result from attempting to load a body of extracted material.

### 7.2.4.4 Ontology Issues

#### *7.2.4.4.1 Occupation Types vs. Position Relations*

An ongoing issue attaches to the representation of what may generally be called "positions". First off, this requires definition, for one of the lessons of the last two years of work is that the word means different things to different people. For the purposes of this document, it will be taken to mean a role that an individual person plays with respect to a particular organization, institution, or project. Obviously, such roles can be grouped into different subtypes, based on a number of different criteria. Some constitute what we ordinarily think of as "jobs" such as "secretary" or "construction worker". Some are very strictly defined in terms of an organizational substructure, such as "logistics officer". Other positions are less formal, less permanent, and more comparatively ad hoc, e.g. "construction foreman". Sometimes there is a noteworthy distinction to be made between an individual's "official" role as indicated by a particular job title or description, and the actual role that the individual plays within the organizational position.

All of these distinctions can be readily taken account of in representation. However, there is also a more fundamental ontological issue having to do just with the question of *how* positions ought to be ontologized. Referring to them as "roles" gives the game away, so far as our preferences are concerned: these things are really two-term *relations* obtaining between persons and the groups or organizations of which they are members. Indeed, one can make the argument that perhaps these should really be thought of as *three*-term relations obtaining between a person, an organization, and the time-frame in which the person serves the organization in that capacity.

The problem here is that resorting to this comparatively rich treatment of positions is often tantamount to presupposing more information than automated extraction is able to provide. This is especially true if the focus is on the extraction of named entities. Often extraction can identify a named entity as a person and link him or her to a particular position, but can do no more, working solely from the source text: e.g., we may know that a particular person is a secretary without being able to tell in what organization he/she holds this position, or for how long. Relying on the two- or three-term relation for translation purposes in such circumstances is potentially disastrous. For this reason, in addition to developing a class of two-term position relations, we have seen fit in IAA-Cyc 2 to also fall back on occasion on the older #$OccupationType vocabulary, which treats positions as CycL collections, e.g., the collection of all persons who are secretaries.

In conclusion, it must be noted that the difficulty that is described here is really only an instantiation of a larger problem applying to any "stative" situational predicate that treats a state or situation involving a set of individuals as a multi-term relation taking those individuals as arguments: if the knowledge base is coupled with an automated knowledge-acquisition facility that can reliably recover some but not all of the individuals involved, we face the frustration of not being able to employ the predicate in assertions for want of terms. The only solution, besides the fairly unpalatable one of introducing lower-arity predicates on an ad hoc basis to deal with the partial knowledge, is to try to develop compositional representations that try to the extent possible to factor the multi-term relation into binary relation ships linking the other players in the situation to some common unifying factor reified for the purpose.

#### *7.2.4.4.2 Expectation Strength and Probability*

The aforementioned impetus to representing different degrees of expectedness as a means toward distinguishing real-world anomaly detection from fact checking is worthy of note in that it appears to assimilate quite closely - perhaps to the point of identity - with one interpretation of probability: i.e., probability as expectation-strength. The extent to which expectation-reasoning overlaps with probability reason thus depends, to some yet-to-be elucidated extent, upon the still-vexed question of how to interpret probabilities. We have not attempted to resolve this question, or do much work on expectation-strength in this phase of the project. We anticipate that the direction of future work with regard to this issue will be determined primarily by the needs and interests of the analysts who would figure as the primary consumers for a projective IAA-Cyc system.

We have not, to date, attempted to implement a system for probabilistic reasoning in Cyc. There have been tentative steps taken towards defining a predicate for specifying degree of expectation for a given assertion, but it has not yet been tested in any extant use case.

# 8 Lessons Learned and Future Directions

### 8.1.1 IAA-Cyc Information Extraction Software Development

The following paragraphs discuss the lessons learned and future directions for the IAA-Cyc IE software development.

**Generality and Scalability.** It is important for the technical approaches used and implemented in the IE software to scale up to real world requirements. They should also be useful and applicable across a wide range of documents, and not just work for a limited set of documents or a "toy" domain. A great deal of data analysis is required to determine which linguistic expressions are instances of infrequent idiosyncratic forms versus common phenomena. In the future, more emphasis will be placed on using data analysis on a wide range of documents to help drive the development of the system.

**Multiple Technical Approaches to Extraction** We learned that no single technology solves the entity identification problem, and that the individual technical approaches have their strengths and weaknesses. For example, the statistical-based approach (e.g., IdentiFinder) has the advantage that it can detect entity names that it has never encountered or seen before. Of course, it will not detect 100% of entity names, and its performance may suffer on text that is not similar to the type of text on which it was trained. So software components using complementary technologies have been incorporated into the IAA-Cyc system including IdentiFinder, a Lexicon Lookup component, and natural language processing components. The purpose of having multiple entity identification components is to improve the recall and precision of the entity identification step over the performance that would be provided by just one entity identification approach.

**Extensibility and Knowledge Acquisition.** Extensibility of system capabilities is required since it is currently impossible for a developer or vendor to provide a "complete" information extraction system that meets all of the needs of a targeted group of end users, let alone a system that will continue to do so in the future. An information system needs to be adaptable as user requirements change, especially as impacted by changes in the real world and in our language. Therefore, knowledge acquisition and extensibility of the IAA-Cyc system's extraction capabilities should be an area of focus in the future. Three types of extensibility will be addressed to varying degrees: extension of capabilities by developers, by end users, and by partially automated means.

**Extend Extraction Capabilities.** The system's extraction capabilities need to be extended to achieve a good useful baseline prototype. IE development work should address extensions to the system so it can extract additional types of attributes, relationships, and meta-information. Extensibility of the three types mentioned in the above paragraph will be addressed and incorporated to the extent possible in the next phase of our development.

**Extend Inference Capabilities.** Extend the system so it can infer additional types of attributes, relationships, and meta-information. This area is important since there is much information to be obtained from a document that is not explicitly stated in the text of a document, but which can be derived (frequently using common sense or real world knowledge).

**Information Comparison and Evaluation.** There is a need for the system to be able to perform information comparison and evaluation to check for (1) general inconsistencies and contradictions, (2) conformity to expectations (expectedness), and (3) relevant anomalies ("the needle in the haystack"). Future directions include development of software capabilities to enable the system to compare and evaluate information that may have been extracted, inferred, and/or acquired from other sources. This information evaluation capability should include the assignment of a measure of confidence to each of the information items.

**Improve Performance.** The performance of the system needs to be improved in the functional areas of information extraction (e.g., recall and precision). These performance areas will be addressed as part of future work.

**Extend User Control and Review Capabilities.** Users have expressed the need for more extensive control and review capabilities. These include capabilities such as controlling which stages of processing are performed, extending or modifying the lexicon, making corrections to items in the IBOK, exporting extracted/derived information to other tools, among others. These capabilities will be targeted in future work.

### 8.1.2 Cyc KB Ontological Engineering

There were three principle lessons learned in the course of the IAA-Cyc 2 work in the area of ontological engineering. First, temporal qualification and intensional inference currently impose fundamental but, we believe, surmountable limitations on anomaly checking. Second, the CycL truth maintenance system imposes a fundamental constraint on conflict checking to the extent that it entails that any form of conflict detection that is implemented in the KB in the near future must have code support and must entail reasoning about error handling. Third, more consultation with bona fide analysts is needed to determine the types of query that are actually of greatest interest and to help resolve several outstanding issues. We deal with these matters in detail in the succeeding section on Future Directions.

The following subsections describe future directions for the ontological engineering work accomplished as part of the IAA-Cyc 2 project. Some of these issues may readily be viewed as implying recommendations scaleable to any approach that seeks to integrate rule-based reasoning with an automated extraction system; others may hold promise for future Cycorp-Veridian collaboration.

#### 8.1.2.1 Resolve Remaining DB-KB Communication Issues

Generally speaking, it seems as if major issues remain concerning facilitation of KB inter-operability with distributed data stores. As previously noted, there is a strong (previously inviolate) assumption in Cyc to the effect that the ground atomic formula assertions on which inference is carried out have the form of CycL assertions and are asserted within reasoning domains in the Cyc knowledge base. There are strong reasons for believing however, that in any application wherein Cyc is expected to interact with an automated extraction system outputting to a distributed database, it will not be practical to have database tuples turned into CycL assertions *en masse* and uploaded to the knowledge base given that an enormous number (hundreds of thousands or millions) of source documents could be involved in a real-world application, implying a database of commensurate size. We believe that this entails a solution according to which any inference work done in relation to the database would be done on delimited sections that were selected and uploaded to the KB for inferencing. How these sections would be selected, and whether the selection was wholely user-driven or system-assisted, are issues that would have to be resolved. The process would very likely be at least partly query-driven, to the extent that the type of query (or queries) to be run would be at least partially determinative of which individual's DB representations were uploaded to Cyc for inferencing.

#### 8.1.2.2 Devise a Query Battery In Consultation With Analysts

A very high priority should initially be placed on simply looking at the output of the Veridian extraction component over a wide range of related text inputs, and trying to determine what types of questions regarding this structured data would be most useful to the analyst. As noted, the reasoning focus has mainly been a kind of proof-of-concept approach designed to try and show a capability for certain types of anomaly- and conformity-checking. However, assuming a suitably wide range of individuals, events, and relationships between them is being extracted, and assuming the existence of a reasonably good translation scheme between the database and Cyc, there is a broad variety of queries that could probably be developed to elucidate relationships of interest to the analyst and the intelligence domain: it simply becomes a question of finding out what these relationships are and queries to get at them. Some of the said queries could likely be accomplished using fairly inexpensive (i.e., code-supported) subsumption reasoning on the Cyc genls and genlPreds hierarchies, or by simple one- or two-backchain inferences that would not be likely to be temporally or computationally expensive to do. We believe this indicates extensive interaction with a group of selected analysts in order to establish what are the main query types of interest and in order to afford analysts means of creating them.

Such discussion can and probably should provide the forum in which the question of whether the anomaly-detection system should be viewed as trapping for real world anomalies or data collection errors is finally resolved.

#### 8.1.2.3 Resolve Issues Concerning Temporal Qualification

An additional reason for including analysts in the discussion would be to help determine precisely what were the temporal qualification issues attaching to which queries. We view this as being largely delimited by analyst interest, and the suspicion is that it could well be the case that the analyst's requirements may be such as would admit approaches that are comparatively more coarse-grained than what we have been attempting to implement. For example, at present it should be regarded as an open question, how important it is that the system actually be able to detect when incommensurate positions are held within a specified time-frame. It might be that this is useful; it might also be that it presupposes information more specific than can readily be extracted from the source document, and it might also be that the mere fact that two "incommensurate" positions were held, absent knowledge about the time frames, was deemed sufficiently interesting to flag the analysts' attention. Such issues can only be resolved with further discussion.

#### 8.1.2.4 Resolve Questions Concerning Inference Engine

Establishing general types and templates for all of the queries to be run against the extracted DB would also have the salubrious effect of helping determine precisely what use should be made of Cyc's internal inferencing facilities. Inexpensive queries and queries that relied on functionality with HL code support could and should be run via AP calls to the inference engine - and of course, if a desired query was not inexpensive or supported in code, this would be an argument for developers to make it so. Generally speaking, this approach would probably be preferable to developing a stand-alone inference facility that relied on specified rules and assertions in the KB.

#### 8.1.2.5 Resolve Intensional Inference Issues re. Expectation Predicates and Implement Conflict-Checking

Finally, more development work is indicated regarding support for the axiomatization of the expectation predicate, and for explicit conflict detection. On the expectation-checking side, Cycorp is near to implementing a supported modal reasoning scheme in which the expectation predicate might be redefined, in a way that would make the intensional inference problem far more tractable. This can only be resolved empirically with support from the ontological engineers and developers involved in the aforesaid project. Similar considerations apply to contradiction checking, with the proviso that here even more direct support from developers is implied. We may take it as axiomatic that any form of outright contradiction-finding that is pursued within IAA or any other context using Cyc will rely primarily on an analysis of error detections deriving from the loading of data to the knowledge base.

# 9 List of Acronyms

| | |
|---|---|
| AFRL | Air Force Research Laboratory |
| API | Application Programmer Interface |
| ASCII | American Standard Code for Information Interchange |
| BOK | Body of Knowledge |
| DIODE | Dynamic Information Operations Decision Environment |
| GAF | Ground Atomic Formula |
| HPKB | High Performance Knowledge Bases |
| IAA | Intelligence Analyst Associate |
| IBOK | Interim Body of Knowledge |
| IE | Information Extraction |
| KB | Knowledge Base |
| KE | Knowledge Engineering |
| NLP | Natural Language Processing |
| OE | Ontological Engineering |
| PLA | People's Liberation Army |
| PLAAF | People's Liberation Army Air Force |
| RKF | Rapid Knowledge Formation |
| RPC | Remote Procedure Call |